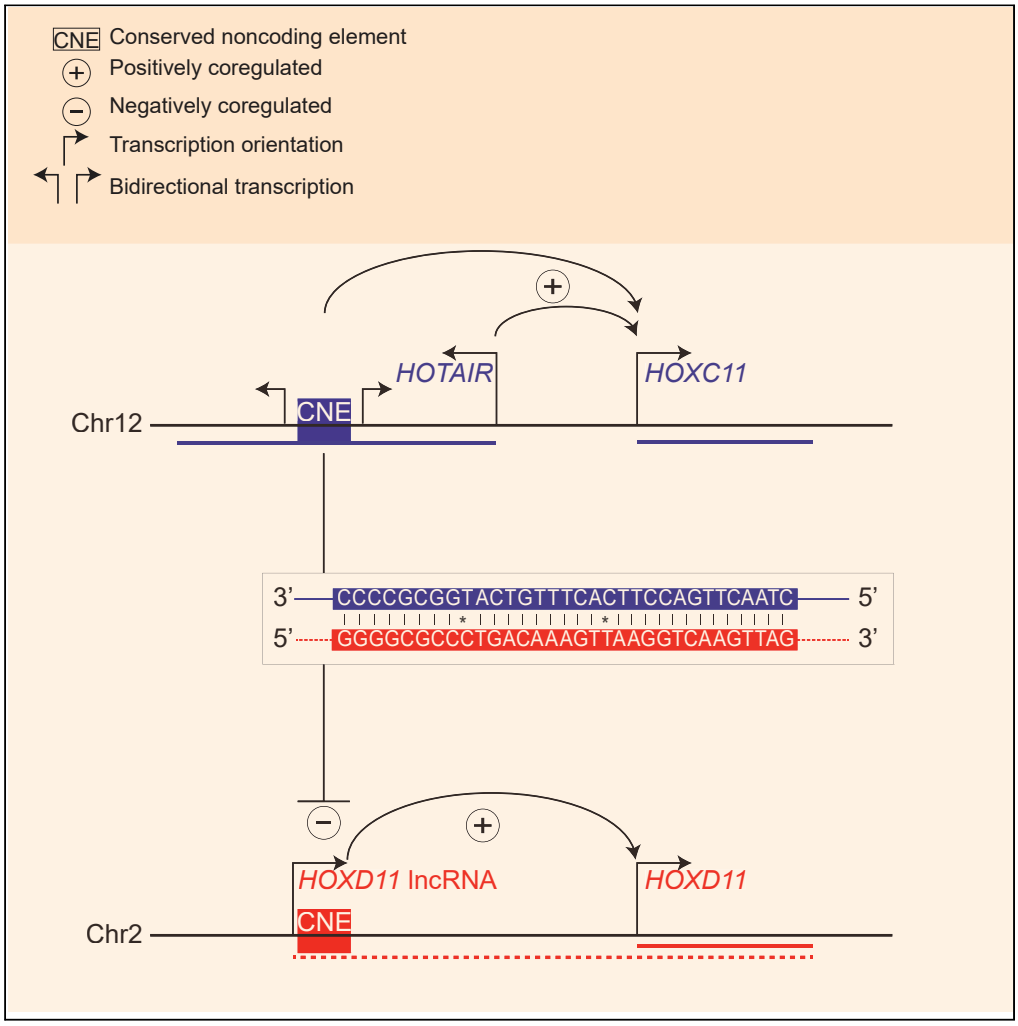
iScience

CellPress
OPEN ACCESS

Article

Ancestrally Duplicated Conserved Noncoding Element Suggests Dual Regulatory Roles of HOTAIR in *cis* and *trans*



Chirag Nepal, Andrzej Taranta, Yavor Hadzhiev, ..., Boris Lenhard, Ferenc Müller, Jesper B. Andersen

chirag.nepal@bric.ku.dk (C.N.)
f.mueller@bham.ac.uk (F.M.)
jesper.andersen@bric.ku.dk (J.B.A.)

HIGHLIGHTS

Two conserved noncoding elements (CNEs) overlap HOTAIR and its paralog on HOXD cluster

Paralog on HOXD cluster may resolve controversy over *cis* and *trans* effects of HOTAIR

CNEs are positively coregulated with *cis* HOX genes but negatively with *trans* cluster

Transcribed CNEs with coevolved complementarity suggest hybridization-based function

Nepal et al., iScience *23*, 101008
April 24, 2020
https://doi.org/10.1016/j.isci.2020.101008

iScience

CellPress
OPEN ACCESS

Article

Ancestrally Duplicated Conserved Noncoding Element Suggests Dual Regulatory Roles of HOTAIR in *cis* and *trans*

Chirag Nepal,[1,7,*] Andrzej Taranta,[1] Yavor Hadzhiev,[2] Sachin Pundhir,[1] Piotr Mydel,[3,4] Boris Lenhard,[5,6] Ferenc Müller,[2,*] and Jesper B. Andersen[1,*]

SUMMARY

***HOTAIR* was proposed to regulate either HoxD cluster genes in *trans* or HoxC cluster genes in *cis*, a mechanism that remains unclear. We have identified a 32-nucleotide conserved noncoding element (CNE) as *HOTAIR* ancient sequence that likely originated at the root of vertebrate. The second round of whole-genome duplication resulted in one copy of the CNE within *HOTAIR* and another copy embedded in noncoding transcript of *HOXD11*. Paralogous CNEs underwent compensatory mutations, exhibit sequence complementarity with respect to transcripts directionality, and have high affinity *in vitro*. The *HOTAIR* CNE resembled a poised enhancer in stem cells and an active enhancer in HOTAIR-expressing cells. *HOTAIR* expression is positively correlated with *HOXC11 in cis* and negatively correlated with *HOXD11 in trans*. We propose a dual modality of *HOTAIR* regulation where transcription of *HOTAIR* and its embedded enhancer regulates *HOXC11 in cis* and sequence complementarity between paralogous CNEs suggests *HOXD11* regulation *in trans*.**

INTRODUCTION

Mammalian genomes are pervasively transcribed, giving rise to thousands of long noncoding RNAs (lncRNAs) (Hon et al., 2017; Iyer et al., 2015). Only a handful of lncRNAs have well-characterized functions, which are attained through diverse mechanisms (chromatin regulation, alternative splicing, gene silencing, *trans*-regulation) (Guttman and Rinn, 2012; Mercer and Mattick, 2013). Although most early studies showed lncRNAs repress gene expression, some lncRNAs have enhancer-like functions and regulate genes in *cis* (Orom et al., 2010). Genomic deletion of lncRNA also removes *cis*-regulatory DNA elements, thus confounding whether the observed phenotype is due to the underlying genomic DNA, the lncRNA transcript itself, or transcription (Bassett et al., 2014; Engreitz et al., 2016; Kaikkonen and Adelman, 2018). As such, transcription blockage and perturbation of the *Lockd* lncRNA showed that it regulates *Cdkn1b* transcription through an embedded enhancer sequence, whereas the lncRNA transcript is dispensable for *Cdkn1b* expression (Paralkar et al., 2016). Deletion of 12 genomic loci encoding various lncRNAs revealed 5 loci whose deletion affected the general process of transcription and enhancer-like activity, but no requirement for the lncRNA transcripts (Engreitz et al., 2016). *Lincp21* locus previously thought to function through its RNA transcript was shown to include multiple enhancers and regulate genes in *cis* (Groff et al., 2016). Moreover, genomic and epigenomic functional annotation have revealed that most intergenic lncRNAs originate from enhancers (Hon et al., 2017). In line with enhancer function overlapping with lncRNAs, the *Haunt* lncRNA has dual roles (Yin et al., 2015), where its DNA encodes enhancers to activate HoxA genes and *Haunt* lncRNA prevents aberrant HoxA expression.

HOTAIR is an intergenic lncRNA located between *HOXC11* and *HOXC12* in chromosome 12. It was proposed to regulate HOXD cluster genes (i.e., *HOXD8*, *HOXD9*, *HOXD10*, and *HOXD11*; located in chromosome 2) in *trans* by recruiting the Polycomb Repressive Complex 2 (PRC2) (Rinn et al., 2007). However, this regulatory model was questioned, as PRC2 binding is promiscuous (Davidovich et al., 2013) and PRC2 was found to be dispensable for *HOTAIR*-mediated transcriptional repression (Portoso et al., 2017). Deletion of the entire Hoxc cluster (including *Hotair*) in mouse showed limited impact on gene expression and H3K27me3 levels at Hoxd genes (Schorderet and Duboule, 2011). Specific deletion of *Hotair* produced a phenotype of homeotic transformation and skeletal malformation as well as genome-wide decrease in

[1]Biotech Research and Innovation Centre, Department of Health and Medical Sciences, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark

[2]Institute of Cancer and Genomics Sciences, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, B15 2TT Birmingham, UK

[3]Broegelmann Research Laboratory, Department of Clinical Science, University of Bergen, Bergen, Norway

[4]Department of Microbiology, Faculty of Biochemistry, Biophysics and Biotechnology, and Malopolska Centre of Biotechnology, Jagiellonian University, Karkow, Poland

[5]Institute of Clinical Sciences MRC Clinical Sciences Centre, Faculty of Medicine, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN, UK

[6]Sars International Centre for Marine Molecular Biology, University of Bergen, 5008, Bergen, Norway

[7]Lead Contact

*Correspondence: chirag.nepal@bric.ku.dk (C.N.), f.mueller@bham.ac.uk (F.M.), jesper.andersen@bric.ku.dk (J.B.A.)

https://doi.org/10.1016/j.isci.2020.101008

H3K27me3 levels and upregulation of posterior HoxD genes (i.e., *Hoxd10*, *Hoxd11*, and *Hoxd13*) (Li et al., 2013). These observations were challenged as specific knockouts of the *Hotair* locus *in vivo* have shown neither homeotic transformation nor upregulation of HoxD genes, but instead a significant change in HoxC (especially *Hoxc11* and *Hoxc12*) cluster genes (Amandio et al., 2016). This strongly argues in favor of a DNA-dependent effect of the *Hotair* deletion (Amandio et al., 2016). Whether *cis*-regulation of *HOTAIR* is mediated via an unannotated enhancer element within its gene body or through transcription of the *HOTAIR* promoter remains unknown. Different regulatory mechanisms (*cis* versus *trans*) might be explained by tissue origin and changes in developmental stages in distinct genetic backgrounds (Li et al., 2016a). As such, there is no consensus model for *HOTAIR*-mediated regulation (Selleri et al., 2016). As the two current models suggest fundamentally different modes of *HOTAIR* function, we decided to revisit the role of *HOTAIR* by a systematic comparative genomic analysis.

To address whether *HOTAIR* regulates HOXC cluster genes in *cis* (Amandio et al., 2016) or HOXD cluster genes in *trans* (Li et al., 2013; Rinn et al., 2007), we exploited comparative sequence analysis across vertebrates and integrated this with transcriptomic and epigenomic data in human and mouse. The HOXC and HOXD clusters originated from an ancestral HOXC/D cluster during the second round of whole-genome duplication (WGD). We hypothesized that the two clusters may contain previously undetected remnants of an ancestral sequence, which might provide important clues on *cis* and/or *trans* interactions. We have identified and characterized a 32-nucleotide conserved noncoding element (CNE) as the *HOTAIR* ancestral sequence, which is shared by both paralogous loci in HoxC and HoxD clusters, presenting itself in an inverted syntenic position. Strikingly, the paralogous CNEs underwent compensatory mutations during vertebrate evolution, which exhibit sequence complementarity dependent on the transcript orientation. Also, the CNEs have high interaction propensity revealed by microscale thermophoresis (MST). These observations suggest direct hybridization between the two noncoding transcripts. *HOTAIR* CNE represents either an active or poised enhancer in different cellular contexts. Its expression is positively correlated with *HOXC11*, whereas negatively correlated with *HOXD11*, suggesting dual modality of *HOTAIR* CNE in *cis* and *trans*.

RESULTS

Identification of *HOTAIR* Ancient Sequence and Its Paralog in HoxD Cluster

The Hox gene clusters are highly conserved across all vertebrates and contain multiple regulatory elements that often have small stretches of highly conserved noncoding elements (CNEs) (Engstrom et al., 2008; Lee et al., 2006). As *HOTAIR* is located within the highly conserved HoxC cluster, we asked whether it has small stretches of conserved sequences that were previously overlooked (He et al., 2011). To this end, we analyzed human and zebrafish annotated CNEs from the synteny analysis tool ANCORA (Engstrom et al., 2008) and identified a 32-nucleotide long CNE (Figure 1A) that is conserved across vertebrates (Figures S1A and S1B). Depending upon the transcript models, the CNE sequence is either located in the intron of Ensembl transcripts or in the exon of an intron-retained alternative transcript annotated in the lncRNA catalog (Figure 1A) (Iyer et al., 2015).

The CNE in zebrafish mapped between *hoxd11a* and *hoxd12a* in the hoxd cluster (Figure 1B), but not in the *hoxC* cluster. To determine whether the CNE is located in the HoxC cluster (the capitalized "HoxC" is used to represent the HoxC cluster across multiple species) or HoxD cluster (Figures S1A and S1B), we systematically mapped CNE sequences across 34 vertebrates and 3 invertebrates (Table S1). Two copies of the CNE were identified in all jawed vertebrates (except in teleosts and birds), but not in the jawless vertebrate lamprey and invertebrates (Figure 1C). The homologous CNEs mapped between *HoxD11* and *HoxD12* (reported target genes of *HOTAIR* in *trans*) in the HoxD cluster and between *HoxC11* and *HoxC12* (reported target genes of *HOTAIR* in *cis*) in the HoxC cluster (Figure 1C) in synteny, suggesting paralogy. The absence of CNE in HoxC cluster in birds might be due to the unassembled HoxC cluster (Table S2). In contrast, teleosts have well-annotated *hoxc11* and *hoxc12* genes in the same cluster (Table S2), but underwent an additional round of teleost-specific WGD, resulting in lineage-specific loss of the paralog. The basal group of jawed vertebrates, such as elephant shark (cartilaginous fish) and spotted gar (basal ray-finned fish; sister group of teleosts) (Figure 1C), have two copies of the CNE suggesting that it was already present in the ancestral HoxC/D cluster and resulted in two copies following the second round of WGD (Figure 1D). Although CNE and its flanking sequences are duplicated from the common ancestral sequence (Figure S1C), the flanking regions have limited homology (for example, in human and elephant shark; Figure S1D). However, CNE and its flanking sequences aligned separately across HoxC and HoxD clusters and revealed a relatively long stretch of sequence conserved across vertebrates (except teleosts) (Figures

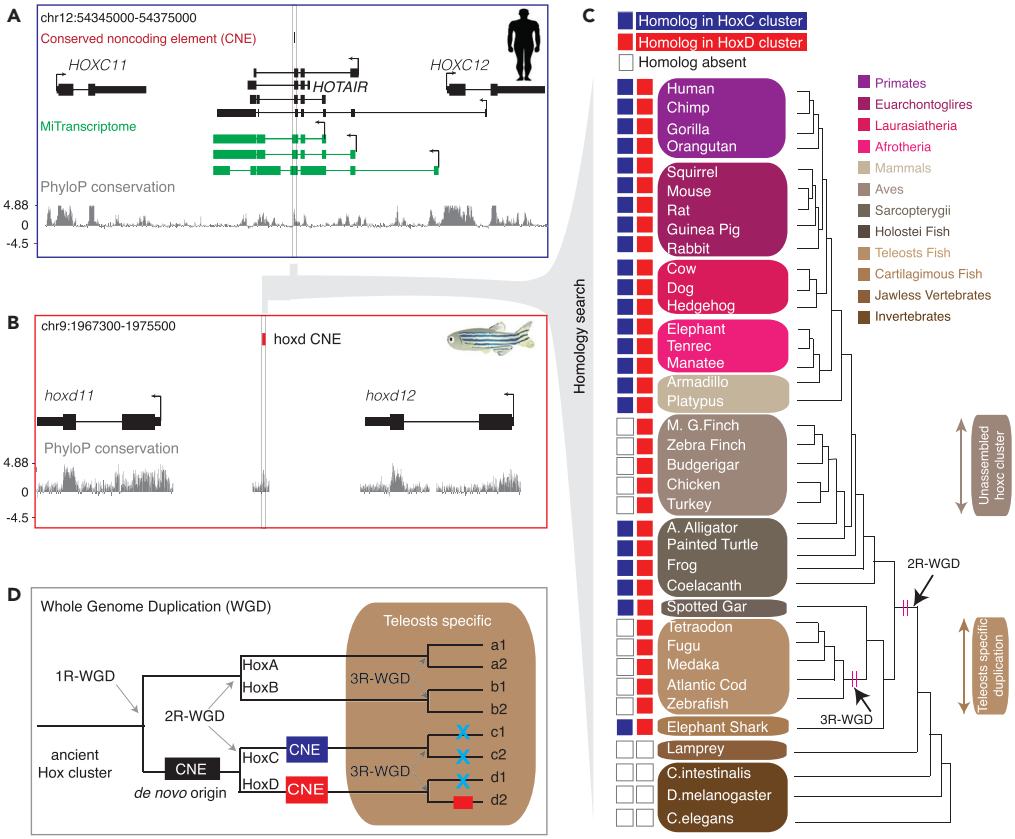


Figure 1. Identification of the *HOTAIR* Conserved Noncoding Element (CNE) and Its Homolog in HOXD Cluster across Vertebrates

(A) A genome browser view around *HOTAIR* locus showing CNE from ANCORA browser and UCSC PhyloP conservation track. The CNE highlighted in a rectangular box is located eight nucleotides away from the splice site.

(B) The ortholog of the CNE mapped to the zebrafish hoxd (between *hoxd11* and *hoxd12*) cluster.

(C) Homology search of the CNE across 37 species identified homologous CNEs in only HoxC and HoxD clusters. Homologs of the CNE are undetected in jawless vertebrate and invertebrates. Homologs in HoxC and HoxD clusters are represented by blue and red, respectively. Empty boxes indicate absence of homologs.

(D) Schematic representation for the proposed model of the origin of the CNE. The CNE might have a *de novo* origin in ancestral HoxC/D cluster where the second round of whole-genome duplication (2R-WGD) resulted two copies in HoxC and HoxD clusters. Teleost-specific duplication might have resulted in loss of CNE from both HoxC clusters and one of the HoxD cluster.

S1E and S1F). Thus we conclude that *HOTAIR* CNE is the ancient sequence and has two paralogous copies in all jawed vertebrates, except teleosts.

Paralogous CNEs Are Transcribed and Embedded in Mature Noncoding Transcripts

Our findings of paralogous CNE in the HoxD cluster suggest the existence of a *HOTAIR* homolog transcript overlying the CNE. To understand whether CNEs are embedded in the mature transcript, we first confirmed that *HOTAIR* CNE can be embedded in the exon by an intron-retained transcript model (Figure 1A). We analyzed long RNA sequencing (RNA-seq) data from ENCODE cell types (Djebali et al., 2012) and observed a large number of reads mapping to introns, particularly the region overlapping CNE, as shown for HeLa S3 cells (Figure 2A). A significant proportion of reads mapped to introns both in whole cell and in nuclear fraction-enriched RNA libraries and was depleted in cytosol-enriched RNA libraries (Figure 2A). We quantified reads mapped to exon, intron, and exon/intron junctions across different cell types and observed a large fraction of reads (relative to exons) mapped to introns in *HOTAIR* (Figure 2B and Table S3), but not in *HOXC11* and *HOXD11* genes (Figure S2A). We observed that a similar pattern of reads mapped to the intron in the region overlapping the CNE across different cell types (Yue et al., 2014), additionally

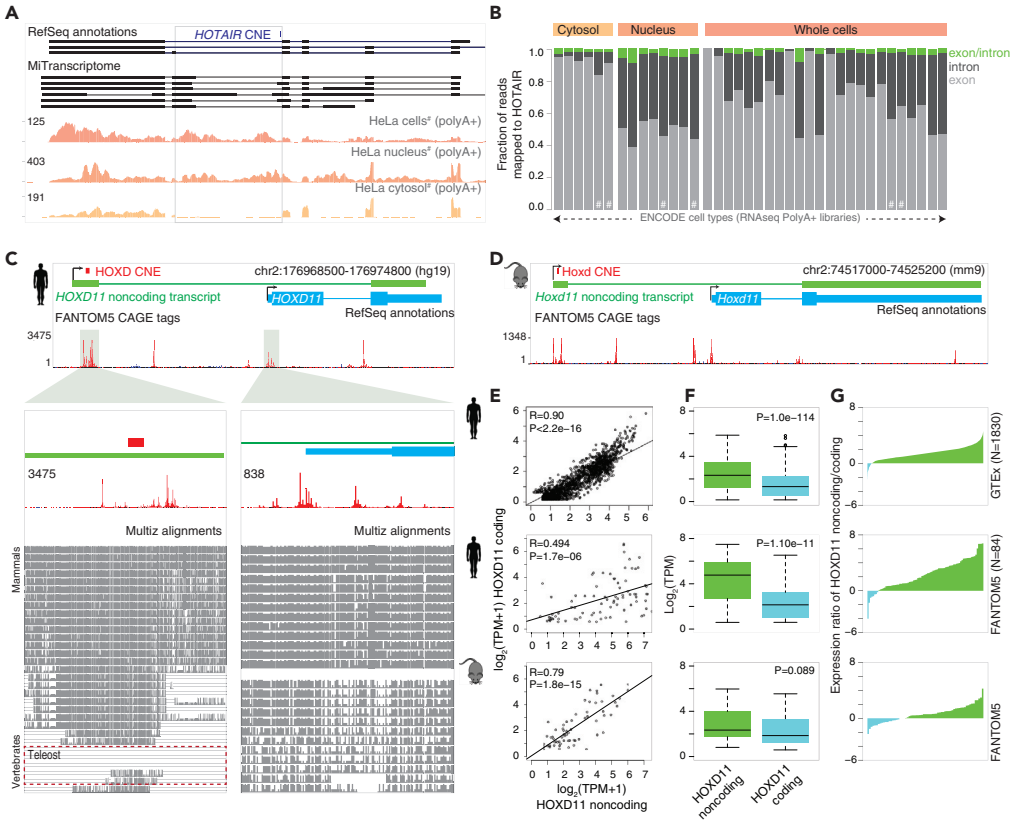


Figure 2. Paralogous CNEs Are Embedded in Mature Noncoding Transcripts

(A) A genome browser view shows HOTAIR transcripts model from RefSeq and MITranscriptome along with RNA-seq coverage tracks from HeLa S3 cells. Large number of reads map to introns in whole cells and nuclear fraction-enriched libraries and are depleted in cytosol fraction-enriched library.

(B) Distribution of reads mapped to *HOTAIR* exon, intron, and overlapping exon/intron junctions across multiple cell types. "#" denotes the HeLa S3 cells. Cell types are ordered based on increasing number of mapped reads.

(C and D) A genome browser view to show transcription and sequence conservation around the *HOXD11* coding (cyan) and *HOXD11* noncoding (*ncHOXD11*) transcripts (green) in human (C) and mouse (D). The zoomed-in promoter regions show lack of sequence conservation of *ncHOXD11*.

(E) Correlation of expression levels of *ncHOXD11* with *HOXD11* coding gene across multiple samples in human (from GTEx and FANTOM5 cohorts) and mouse (FANTOM5 cohort).

(F) Expression levels of *ncHOXD11* and *HOXD11* coding gene across multiple samples from GTEx and FANTOM.

(G) Ratio of expression levels of *ncHOXD11* and coding gene across individual cell types. Positive value on y axis indicates higher expression levels of *ncHOXD11*.

confirming that intron retention of *Hotair* is conserved in mouse (Figures S2B and S2C). Collectively, this suggests that the *HOTAIR* CNE is embedded in an intron-retained transcript.

To associate HOXD CNE with the transcript models, we intersected it with Ensembl transcripts and identified that the CNE is embedded in the exon of a previously annotated noncoding transcript sharing the locus with *HOXD11* coding gene in human (Figure 2C) and mouse (Figure 2D). The CNE is located in the promoter region of *HOXD11* noncoding transcript (referred as *ncHOXD11* from here on), which is approximately 55 nucleotides downstream of the dominant transcription start site (denoted by the highest CAGE peak in FANTOM5 data) in human and mouse (Figures 2C and 2D). Given the shared locus, we sought to understand the nature and extent of *ncHOXD11* usage in relation to the *HOXD11* coding gene. The expression level of *ncHOXD11* is positively correlated with *HOXD11* coding gene across GTEx (GTEx Consortium, 2013) and FANTOM5 (Arner et al., 2015; FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014) data in human and mouse (Figure 2E). The expression of *ncHOXD11* is significantly higher than that of the *HOXD11* coding gene (Figure 2F) in the majority of cell types in human (Figure 2G). However, in mouse, the expression of *ncHoxd11* is only marginally higher than that of the *Hoxd11* coding gene (Figures 2F and

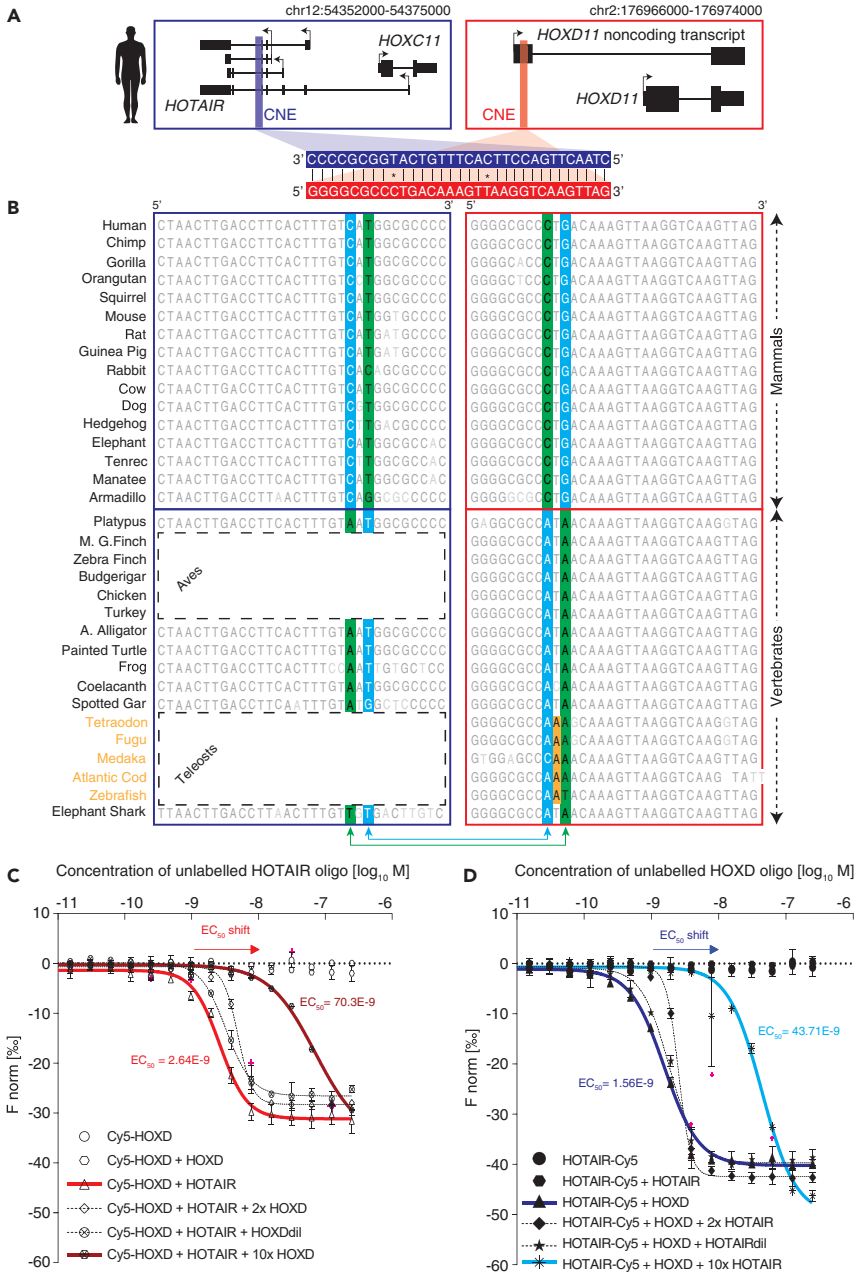


Figure 3. Paralogous CNEs Exhibit Sequence Complementarity in Transcript Orientation

(A) Paralogous *HOTAIR* CNE (blue bar) and HOXD CNE (red bar) are zoomed and aligned in 5′ to 3′ orientation of respective transcripts.

(B) Alignment of the paralogous CNEs in 5′ to 3′ orientation of respective transcripts reveals sequence complementarity across vertebrates. Genetic substitutions within paralogous CNEs co-occurred at specific positions, which resulted in gain or loss of complementarity, where green represents non-complementary DNA and cyan represents complementary DNA. Teleost-specific change in DNA sequence is shown in orange.

(C and D) Microscale thermophoresis (MST) assay to evaluate the interaction between labeled and unlabeled RNA-oligos at different concentration. MST-on time of 5 s was used for analysis. Baseline-corrected normalized fluoresce (ΔF_{Norm}) was chosen to present data (independent n $\geq$ 3 measurements; each point on the graphs presents mean $\pm$ SD). An extrapolated EC_{50} $\pm$ SD curve is fitted and shown on the graph. The concentration of the labeled RNA-oligo was constant at 5 nM. The concentration of unlabeled RNA-oligo was varied at 250 nM to 7.63 pM. The x axis represents the concentration of titrated unlabeled RNA-oligo. The y axis represents interaction-driven normalized fluorescence change

Figure 3. *Continued*
(ΔFnorm[‰]). Measurement of interaction between (C) labeled HOXD CNE (Cy5-HOXD) RNA-oligo and unlabeled HOTAIR CNE RNA-oligo and (D) labeled HOTAIR CNE (HOTAIR-Cy5) RNA-oligo and unlabeled HOXD CNE RNA-oligo.

2G), suggesting a relative gain in the expression of *ncHOXD11* in human. Finally, we analyzed RNA-seq transcript models across species (Basu et al., 2016; Hezroni et al., 2015) and detected both transcripts in ferret and dog, whereas only *ncHoxD11* in chicken (Figure S2D). The location of HoxD CNE, which is downstream of *ncHoxD11* start site, is conserved across species, suggesting that its transcription from *ncHoxD11* is an ancient phenomenon. However, *ncHoxD11* was undetected in teleosts (zebrafish and tetraodon; Figure S2D), which is further supported by absence of *ncHoxD11* promoter sequence (Figure 2A). Collectively, we showed that paralogous CNEs are transcribed and embedded in mature transcripts across multiple species.

Transcribed CNEs Exhibit Conserved Sequence Complementarity across Vertebrates

As paralogous CNEs are embedded in mature transcripts, we sought to analyze their directionality with respect to transcript orientation. We observed that human CNEs exhibit sequence complementarity in transcript orientation (Figure 3A). To ensure that the observed sequence complementarity in human is not by chance we analyzed its orientation in other vertebrates. As transcriptional evidence of *HOTAIR* and *ncHOXD11* was limited to a subset of species (Figure S2E), we inferred orientation for missing transcripts (see Methods), as illustrated for chimp and painted turtle (Figures S3A and S3B). We then aligned CNEs in the transcript orientation and observed sequence complementarity across vertebrates (Figures 3B and S3C). This suggests that sequence complementarity between CNEs is an ancient feature that has been under selection pressure for more than 300 million years. It raises an important question as to whether the key function of these transcripts is to provide transcription of the CNE.

Interestingly, in addition to the conservation and retention of sequence complementarity, we observed that paralogous CNEs revealed a specific pattern of genetic substitution at two specific positions in both CNEs that co-evolved in two separate waves in vertebrates and mammals (Figure 3B). The sequence pairs that co-evolved at two specific positions are depicted in green (non-complementary) and cyan (complementary) (Figure 3B). The nucleotides "A" colored by green in vertebrates are non-complementary, where both nucleotides co-evolved simultaneously in mammalian lineage resulting in gain of complementarity (highlighted by cyan). On the other hand, nucleotides "A" and "T" highlighted by cyan in vertebrates are complementary, where the nucleotide "A" evolved in mammals resulting in loss of complementarity. In mammals, one substitution resulted in retention of complementarity and the other substitution resulted in loss of complementarity, reflecting that paralogous CNEs underwent compensatory mutations. Unlike vertebrates, the HoxD CNE in teleosts evolved separately in its own lineage (highlighted in orange) reflecting no selection pressure to retain sequence complementarity as its putative binding partner in HoxC cluster is lost. Collectively, the coevolution of CNEs and retention of sequence complementarity in the transcript orientation raises the potential for such hybridization based on *trans* function.

Hybridization of Paralogous CNEs *In Vitro*

To verify whether paralogous CNE transcripts hybridize, we designed two Cy5-labeled RNA-oligos (Table S4) for HOXD CNE (Cy5-HOXD) and HOTAIR CNE (HOTAIR-Cy5) and analyzed the interaction propensity using MST (Asmari et al., 2018; Duhr and Braun, 2006a, b; Moon et al., 2018). For labeled Cy5-HOXD RNA-oligo (5 nM), we analyzed the binding with unlabeled *HOTAIR* CNE RNA-oligo titrated at concentrations ranging between 250 nM and 7.63 pM. Similarly, for labeled HOTAIR-Cy5 RNA-oligo, we analyzed the binding with titrated unlabeled HOXD CNE RNA-oligo (Table S5 and Figures S4A–S4C). The labeled Cy5-HOXD and unlabeled *HOTAIR* CNE RNA-oligo showed a strong interaction at the nanomolar scale (EC_{50} = 2.64 × 10^{-9}) (Figure 2C; red line), whereas we observed no binding at control conditions (either labeled oligo alone or mix of labeled and unlabeled counterparts). Similarly, the labeled HOTAIR-Cy5 and unlabeled HOXD CNE RNA-oligo showed a strong interaction at the nanomolar scale (EC_{50} = 1.56 × 10^{-9}) (Figure 2D; blue line), whereas the control showed no binding. To evaluate if an unlabeled oligo can affect the interaction between labeled Cy5-HOXD RNA-oligo and the unlabeled HOTAIR CNE RNA-oligo, we added unlabeled HOXD RNA-oligo and observed that the interaction (EC_{50} = 70.3 × 10^{-9}) was sensitive to the presence of unlabeled RNA-oligo (Figure 3C; dark red line). A 10-fold excess of the competitor resulted in a shift of the fluorescent signal resembling depletion of the titrated oligos and correspondingly a shift

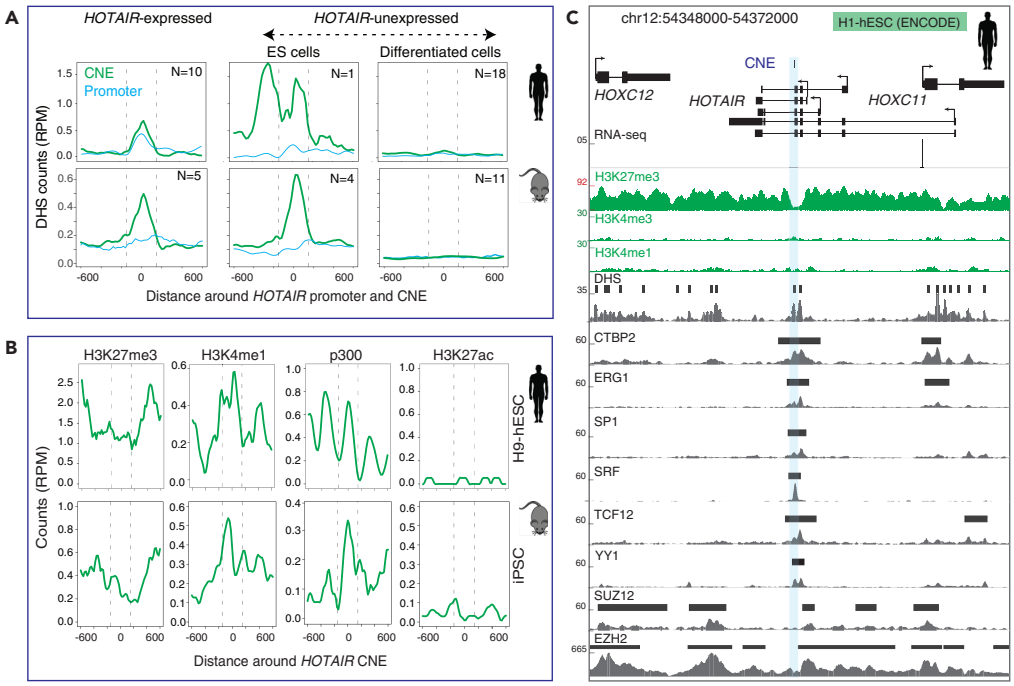


Figure 4. The *HOTAIR* CNE Represents a Poised Enhancer in *HOTAIR*-Nonexpressing Stem Cells in Human and Mouse

(A) Average DNase I hypersensitive site (DHS) signals around *HOTAIR* CNE in *HOTAIR*-expressing and *HOTAIR*-nonexpressing cells (embryonic stem cells and differentiated cells). The y axis is normalized DHS coverage in reads per million (RPM). "N" denotes the number of cell lines.

(B) The distribution of H3K4me1, H3K27me3, H3K27ac, and p300 signals around *HOTAIR* CNE in H9-hESC (human) and iPSC (mouse) cell lines. The y axis is normalized coverage in reads per million (RPM).

(C) A genome browser view with transcription factors, DHS, histone modifications, and RNA-seq tracks from H1-hESC cell line. *HOTAIR* is not expressed in H1-hESC (as shown by lack of RNA-seq reads) and marked by broad H3K27me3 peak. H3K27me3 signal is depleted around CNE, reflecting a nucleosome-depleted region and bound by multiple transcription factors.

in EC_{50} value (Figure 3C; dark red line). Similarly, addition of an unlabeled *HOTAIR* CNE RNA-oligo affected the interaction of the labeled HOTAIR-Cy5 RNA-oligo and unlabeled HOXD CNE RNA-oligo, resulting in a shift in fluorescent signal (Figure 3D; cyan line). Even at low concentration of the competitor oligo the shift is still clear, confirming that the paralogous CNEs have strong interaction *in vitro*.

Chromatin Structure of *HOTAIR* CNE Represents a Poised Enhancer in Stem Cells

CNEs are putative *cis*-regulatory elements (Bejerano et al., 2004; Harmston et al., 2013; Sandelin et al., 2004), and many of them have been experimentally validated as tissue-specific enhancers (Nobrega et al., 2003; Pennacchio et al., 2006; Woolfe et al., 2005). We analyzed experimentally validated enhancers (Pennacchio et al., 2006) and found that the genomic regions overlapping CNEs were not probed for enhancer activity. However, in the literature, we found that the region overlapping the Hoxd CNE was tested for enhancer activity in mouse and shown to drive expression in a proximal posterior part of the developing forelimbs (Beckers et al., 1996). However, subsequent deletion of Hoxd CNE revealed no phenotype *in vivo* (Beckers and Duboule, 1998) (see Discussion).

To understand whether chromatin states of CNEs resemble that of enhancers (Andersson et al., 2014; Pundhir et al., 2016; Roadmap Epigenomics Consortium et al., 2015), we selected 29 cell lines (Table S6) from Roadmap Epigenome project. Based on the expression levels of *HOTAIR* (see Methods), cell lines were classified into HOTAIR-expressing (N = 10) and HOTAIR-non-expressing (N = 19) groups (Figure S5A). The H1-hESC cell line is unique as it is enriched for H3K27me3 and DNase hypersensitive sites (DHSs) (Figure S5B), thus we separately analyzed H1-hESC cells from remaining *HOTAIR*-non-expressing cells. The

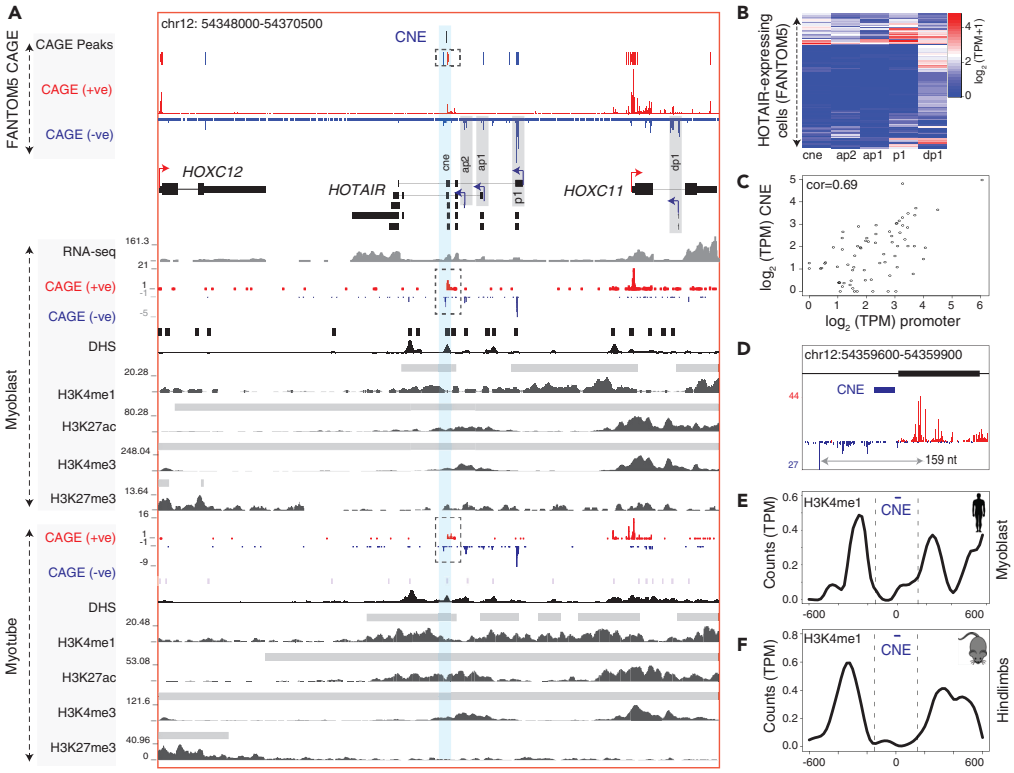


Figure 5. The CNE Represents an Active Enhancer RNA in *HOTAIR*-Expressing Cells

(A) A genome browser view with FANTOM5 CAGE tags (combined tracks) and individual tracks on myoblast and myotube along with RNA-seq and histone modifications. Horizontal bars above histone and DHS tracks are annotated peaks. CAGE tags on forward and reverse strand are represented by blue and red, respectively. Bidirectional CAGE tags flanking the CNE are shown in dashed rectangular boxes. Bidirectional CAGE tags overlap with H3K4me1 and H3K27ac peaks in myoblast and myotube.

(B) Expression levels of the *HOTAIR* CNE and alternative promoters across FANTOM5 samples.

(C) Correlation of the expression levels of the CNE with that of *HOTAIR* promoter.

(D) Bidirectional transcription (from myoblast and myotube differentiation time points) around the CNE roughly represents the length of a nucleosome.

(E and F) Bimodal H3K4me1 peaks flank the CNE in human myoblast (E) and embryonic day 10.5 hindlimbs in mouse (F).

HOTAIR CNE has open chromatin in HOTAIR-expressing cells and HOTAIR-non-expressing stem cells and closed chromatin in HOTAIR-non-expressing differentiated cells in both human and mouse (Figure 4A and Table S6). We observed a similar chromatin state around HOXD CNE (Figure S5C). The chromatin state of CNE is dynamically regulated during reprogramming of mouse embryonic fibroblasts to induced pluripotent stem cells (iPSCs) (Chronis et al., 2017) where the CNE has closed chromatin in mouse embryonic fibroblasts and open chromatin in iPSC (Figure S5D). In addition, enrichment of H3K4me1, H3K27me3, and p300 signals at the CNE in human H9-hESC and mouse iPSC (Figure 4B) provides evidence that the *HOTAIR* CNE represents an embryonic stem cell-specific poised enhancer (Rada-Iglesias et al., 2011). This is further supported by enrichment of enhancer-associated transcription factors (Sigova et al., 2015), such as CTBP2, CHD1, SP1, and YY1 that are exclusively enriched at CNE in H1-hESC (Figure 4C and Table S7). However, p300, H3K4me1, and bimodal H3K27me3 peaks were not enriched around HOXD CNE in hESC (Figures S5E–S5G). As HOXD CNE overlaps with the promoter of *ncHOXD11* (Figures 2A and S5G), genomic analyses of chromatin states will be unable to distinguish a putative enhancer from an overlapping promoter. Collectively, these data suggest that *HOTAIR* CNE resembles a poised enhancer in stem cells in both human and mouse.

HOTAIR CNE Represents an Active Enhancer RNA

To determine whether the CNE represents an active enhancer in HOTAIR-expressing cells, we analyzed FANTOM5 CAGE data (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014) to identify

unstable bidirectional transcription, a hallmark of active enhancer RNA (eRNA) (Andersson et al., 2014). We observed bidirectional transcription flanking the CNE (Figure 5A; dashed rectangular box), providing evidence of an active eRNA. Importantly, transcription of *HOTAIR* primary and alternative promoters is generally co-expressed with bidirectional transcription around the CNE (Figure 5B and Table S8). This is exemplified during myoblast to myotube differentiation (Figure S6A), which suggests coregulation. The expression of CNE is positively correlated with expression from the *HOTAIR* promoter and alternative promoters (Figure 5C), with the exception of the distal promoter (labeled as dp1) (Figure S6B). Negative correlation of the distal promoter is mostly due to a majority of samples in which only the distal promoter is expressed (Figure 5B). The distance between bidirectional transcription start sites flanking the CNE is about the length of one nucleosome (Figure 5D), which is conserved across vertebrates (Figure S1E), and suggests that the CNE shares an evolutionary conserved typical enhancer structure. On the contrary, no evidence suggests bidirectional transcription around the HOXD CNE, as it overlaps with the promoter region of *ncHOXD11* (Figure 2C).

Next, we focused on myoblast and myotube cell types, for which RNA-seq, histone modifications, and DHS data are available as complements to CAGE tags. The RNA-seq reads mapped on introns and across intron/exon boundary around the CNE (Figure 5A), thus providing evidence for an intron-retained transcript. Furthermore, DHS, H3K4me1, and H3K27ac peaks are enriched around the CNE (Figure 5A) in myoblast and myotube, providing additional evidence for an active enhancer. The observed bimodal H3K4me1 peaks around CNE are a characteristic feature of active enhancers (Figure 5E). However, in mouse, relevant tissues and stages wherein *Hotair* is expressed (Amandio et al., 2016; Li et al., 2013; Schorderet and Duboule, 2011) were not included in FANTOM5 samples (Figure S6C) and lack CAGE tags around the CNE. However, H3K4me1 and H3K27ac are enriched in mouse embryonic day 10.5 hindlimbs (Andrey et al., 2017) (Figures 5F and S6D) around the CNE. Thus, we showed that *HOTAIR* CNE resembles an active enhancer in *HOTAIR*-expressed cells, in both human and mouse. Importantly, transcription of the *HOTAIR* promoter is tightly linked to enhancer activity of the CNE, suggesting that its transcription might be a contributor to the purifying selection acting on the CNE and further provides support to the notion that the CNE acts as a regulator in *cis* as previously proposed (Amandio et al., 2016).

HOTAIR Expression Highlights Simultaneous Regulation of Known Target Genes in *cis* and *trans*

As we showed that transcribed CNEs exhibit sequence complementarity in transcript orientation, we sought to understand whether *HOTAIR* can simultaneously regulate genes in *cis* and *trans* mediated via the CNE. We analyzed transcription levels of the *HOTAIR* enhancer with HOX clusters genes across 694 cell types from FANTOM5. The *HOTAIR* was expressed in 104 cell types (Figure S7A), and HOX genes were positively correlated with other genes in the cluster (Figure S7B). The expression of *HOTAIR* CNE with HOXC/D clusters posterior genes on 104 cell types revealed positive correlation with HOXC cluster genes and a trend toward negative correlation with HOXD cluster genes (Figure S7C). To ensure that the correlations were not driven by missing expression of HOX genes, we reanalyzed data by including only those cell types wherein both *HOTAIR* and HOX genes are coexpressed. We observed similar correlations wherein HOXC cluster genes are positively correlated and HOXD cluster genes are negatively correlated (Figure 6A). Strikingly, *HOXC11* is the most positively correlated ($R = 0.60$; p value: 2.6×10^{-9}) and *ncHOXD11* ($R = -0.32$; p value: 0.009) and coding transcripts ($R = -0.29$; p value: 0.02) are the most negatively correlated, both of which are previously reported target genes (Amandio et al., 2016; Li et al., 2013; Rinn et al., 2007). This observation was further validated in 2,436 tissue samples from GTEx and 605 patients with breast cancer (see Methods) from The Cancer Genome Atlas (Pereira et al., 2016) where *HOXC11* had the most significant positive correlation and *HOXD11* had the most significant negative correlation (Figures 6B and S7D).

Therefore, we propose a model to explain the observed correlation between *HOTAIR* expression and that of *HOXC11* and *HOXD11*, a dual regulatory mechanism mediated via the CNE sequence. The positive correlation between *HOTAIR* and *HOXC11* might be mediated via an active eRNA, the act of transcription of *HOTAIR*, or the combination of both. We observed positive correlation between *ncHOXD11* promoter encoding HOXD CNE and *HOXD11* coding gene. Transcriptional activity of the CNE is coupled to *HOTAIR* transcription, suggesting that a key function of the *HOTAIR* transcript could be to provide active transcription for the CNE. Paralogous CNEs embedded in intron-retained *HOTAIR* and *ncHOXD11* transcripts have retained sequence complementarity in transcript orientation that might facilitate hybridization between two RNA transcripts. This hybridization between two RNA transcripts downregulates *HOTAIR* target

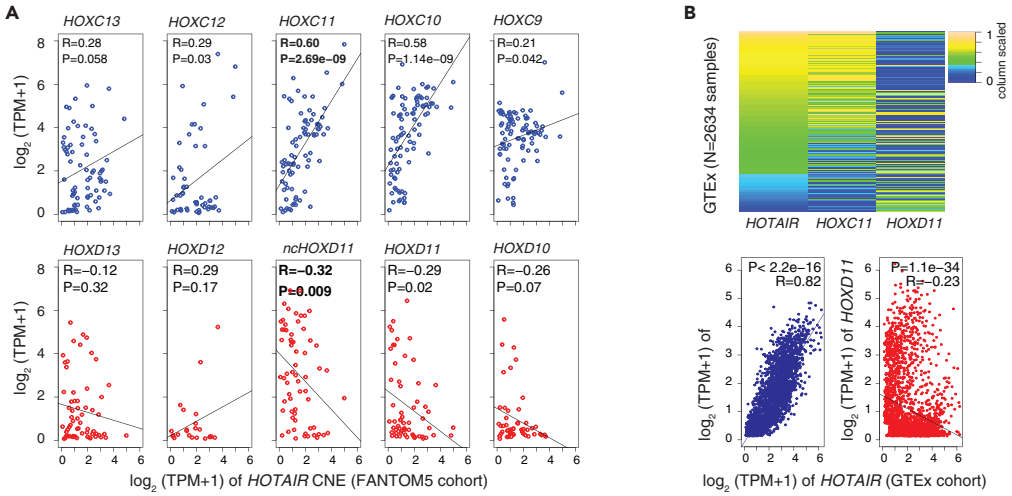


Figure 6. Coregulation of *HOTAIR* with Validated Target Genes *HOXC11* and *HOXD11*

(A) Correlation of expression levels of *HOTAIR* CNE with HOXC and HOXD cluster posterior genes across FANTOM5 cell types. The x axis represents expression level of *HOTAIR* CNE, and the y axis represents the expression levels of HOXC and HOXD cluster genes. Expression level is measured as tags per million (TPM). The expression levels of *HOXC11* have the highest positive correlation, and those of the *HOXD11* coding and noncoding have the highest negative correlation.

(B) Heatmap and correlation of expression levels of *HOTAIR* with *HOXC11* and *HOXD11* genes across GTEx cohort.

gene expression. Thus, we propose that *HOTAIR* can have dual regulatory roles in *cis* and *trans*, which is likely mediated by the CNE paralog sequence.

DISCUSSION

We have identified and characterized a 32-nucleotide CNE as the ancestral sequence that probably originated in ancestral HoxC/D cluster, where the second round of WGD gave rise to one copy in the *HOTAIR* locus and another copy in the *ncHOXD11* locus. The paralogous CNEs are only 32 nucleotides, whereas the conserved sequence flanking the *HOTAIR* CNE is much longer (Figure S1E) and coincides with a region of eRNA, suggesting an ancestral sequence within the *HOTAIR* locus. The remainder of the *HOTAIR* sequence has limited homology in vertebrates (Figures S1B and S1E), which evolved rapidly in mammalian lineage (He et al., 2011). This could be indicative of the *HOTAIR* locus originating from the CNE and evolution favoring the development of its sequence, likely expanding its functionality. Although thousands of CNEs are annotated, only a small minority of them have retained a duplicated copy (McEwen et al., 2006). As such, retention of both copies of *HOTAIR* CNE had not been reported before. To the best of our knowledge, this is the first instance of reported paralogous CNEs that underwent compensatory mutation and have retained sequence complementarity in their transcribed directionality (Figures 3A and 3B).

Many of the experimentally tested CNEs are validated enhancers (Nobrega et al., 2003; Pennacchio et al., 2006; Woolfe et al., 2005). Genome-wide transcriptomic and epigenomic analyses revealed that enhancers are characterized by distinct transcription and chromatin states (reviewed in Li et al., 2016b), and we used these features to define whether CNEs are enhancers. The *HOTAIR* CNE region is marked by open chromatin that is flanked by enriched H3K4me1 and H3K27ac peaks along with bidirectional transcription, which collectively meets all characteristic features of an active enhancer. On the other hand, our genomic analyses did not reveal any enhancer features on HOXD CNE, likely because it overlaps with the *ncHOXD11* promoter region (Figures 2C and 2D), and it is therefore difficult to entangle overlapping signals. However, the sequence overlapping Hoxd CNE drives expression in a proximal posterior part of the developing forelimbs in mouse (Beckers et al., 1996). Recent findings suggest that some promoters have dual functions as enhancers and influence the expression of a neighboring gene in *cis* (Engreitz et al., 2016; Paralkar et al., 2016; Yin et al., 2015). Thus, it is plausible that the *ncHOXD11* promoter overlapping HOXD CNE has enhancer function and regulates *HOXD11* gene in *cis*.

Multiple enhancers with similar activity provide an effective buffer to prevent deleterious phenotypic consequences upon loss of individual enhancers (Osterwalder et al., 2018). As the *HOTAIR* CNE has a

paralogous copy, how this might affect *HOTAIR* regulation needs further consideration. Deletion of a sequence overlapping Hoxd CNE revealed no phenotype *in vivo* (Beckers and Duboule, 1998), which was different from *in vitro* (Beckers et al., 1996). It was speculated that the difference(s) might be due to other phenotypes that were undetected or might have a redundant copy that masked the effect. In fact, we now have identified that the probed sequence has a paralogous copy in the *Hotair* locus that might have masked the effect *in vivo*. Thus, whether paralogous CNEs have redundant functions, such that deletion of one CNE might be compensated by the other, remains unclear. Putting this in the context of deletion of the *Hotair* locus *in vivo* (Amandio et al., 2016; Li et al., 2013), it remains unknown whether the effects of *Hotair* CNE deletion are compensated for, to a certain extent, by paralogous Hoxd CNE.

Transcriptional activity of CNEs is coupled to *HOTAIR* and *ncHOXD11* transcription, suggesting that a key function of these transcripts is to provide active transcription for the CNEs. With respect to transcript orientation, paralogous CNEs exhibit sequence complementarity, which raises the potential for this hybridization principle based on *trans* function. This is supported by the observed hybridization *in vitro* (Figures 3C and 3D) and needs future experiments to confirm *in vivo*. Transcription of *HOTAIR* CNE is positively correlated with *HOXC11* (Figure 6), and transcription of *ncHOXD11* is positively correlated with *HOXD11*. Simultaneously, the transcription of *HOTAIR* CNE and *ncHOXD11* are negatively correlated (Figure 6), which is likely mediated via sequence complementarity between CNEs.

In summary, our analyses suggest that *HOTAIR* could regulate both HoxC and HoxD cluster genes simultaneously and provide a unifying model of *HOTAIR* regulation that should clarify ongoing controversies (Amandio et al., 2016; Li et al., 2013; Portoso et al., 2017; Rinn et al., 2007; Schorderet and Duboule, 2011). Our work highlights how an lncRNA locus could possibly function at the DNA and RNA levels to regulate genes both in *cis* and *trans*. Unraveling such lncRNAs and determining/validating mechanisms through which they function at the DNA and/or RNA levels is an ongoing challenge. We propose that such integrative analyses bridging evolutionary genomics and comparative transcriptomics/epigenomics could prove a powerful tool for better understanding of lncRNA-dependent regulation processes.

Limitations of the Study

Our conclusion is based on analyses of large-scale genomics data, thus future work is needed to validate the predicted models *in vivo*. We showed hybridization between paralogous CNEs *in vitro*, which needs to be validated *in vivo*. Furthermore, targeted experiments are required to understand how specific deletion of individual CNE(s) along with simultaneous deletion of both CNEs alters *HOTAIR*-dependent regulation in *cis* and *trans*.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

Data used in the study were downloaded from ENCODE, mouse ENCODE, GTEx, TCGA, FANTOM5, NIH Roadmap Epigenome project, and additional publicly available datasets mentioned in the methods. All custom code is available upon request.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101008.

ACKNOWLEDGMENTS

We are very grateful to the ENCODE, FANTOM5, GTEx, Roadmap Epigenome consortia, and researchers for making data freely available. Also, the results here are in parts based on data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The authors thank Dr. Colm J. O'Rourke, Dr. Michal Lubas, and Dr. Albin Sandelin for critical comments on the manuscript and Dr. Juan Francisco Lafuente Barquero for help with Illustrator and fruitful discussion. C.N. and A.T. are recipients of a postdoctoral fellowship from the Danish Medical Research Council (6110-00557A) and Lundbeck Foundation (R219-2016-718), respectively. The laboratory of J.B.A. is supported by the Danish Medical Research Council (4183-00118A), Danish Cancer Society (R98-A6446) and Novo Nordisk Foundation (14040). P.M. acknowledges funding from National Science Center (2014/14/E/NZ6/00162, Poland). F.M. and B.L. thank the support of BBSRC (BB/L010488/1) and the Wellcome Trust Investigator Award (106955/Z/15/Z).

AUTHOR CONTRIBUTIONS

C.N. conceived the story. C.N., A.T., Y.H., and S.P. analyzed data. C.N., A.T., B.L., F.M., and J.B.A. interpreted the results. A.T., Y.H., S.P., P.M., A.T., B.L., F.M., and J.B.A. contributed to critical discussions. C.N., F.M., B.L., and J.B.A. drafted the manuscript, and all authors contributed to revising the manuscript.

DECLARATION OF INTERESTS

Authors declare no conflict of interest and no competing financial interest.

Received: November 20, 2019
Revised: February 6, 2020
Accepted: March 18, 2020
Published: April 24, 2020

REFERENCES

Amandio, A.R., Necsulea, A., Joye, E., Mascrez, B., and Duboule, D. (2016). Hotair is dispensible for mouse development. PLoS Genet. *12*, e1006232.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Andrey, G., Schopflin, R., Jerkovic, I., Heinrich, V., Ibrahim, D.M., Paliou, C., Hochradel, M., Timmermann, B., Haas, S., Vingron, M., et al. (2017). Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. Genome Res. *27*, 223–233.

Arner, E., Daub, C.O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drablos, F., Lennartsson, A., Ronnerblad, M., Hrydziuszko, O., Vitezic, M., et al. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science *347*, 1010–1014.

Asmari, M., Ratih, R., Alhazmi, H.A., and El Deeb, S. (2018). Thermophoresis for characterizing biomolecular interaction. Methods *146*, 107–119.

Bassett, A.R., Akhtar, A., Barlow, D.P., Bird, A.P., Brockdorff, N., Duboule, D., Ephrussi, A., Ferguson-Smith, A.C., Gingeras, T.R., Haerty, W., et al. (2014). Considerations when investigating lncRNA function in vivo. Elife *3*, e03058.

Basu, S., Hadzhiev, Y., Petrosino, G., Nepal, C., Gehrig, J., Armant, O., Ferg, M., Strahle, U., Sanges, R., and Muller, F. (2016). The Tetraodon nigroviridis reference transcriptome: developmental transition, length retention and microsynteny of long non-coding RNAs in a compact vertebrate genome. Sci. Rep. *6*, 33210.

Beckers, J., and Duboule, D. (1998). Genetic analysis of a conserved sequence in the HoxD complex: regulatory redundancy or limitations of the transgenic approach? Dev. Dyn. *213*, 1–11.

Beckers, J., Gerard, M., and Duboule, D. (1996). Transgenic analysis of a potential Hoxd-11 limb regulatory element present in tetrapods and fish. Dev. Biol. *180*, 543–553.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321–1325.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative binding of transcription factors orchestrates reprogramming. Cell *168*, 442–459 e420.

Davidovich, C., Zheng, L., Goodrich, K.J., and Cech, T.R. (2013). Promiscuous RNA binding by Polycomb repressive complex 2. Nat. Struct. Mol. Biol. *20*, 1250–1257.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Duhr, S., and Braun, D. (2006a). Optothermal molecule trapping by opposing fluid flow with thermophoretic drift. Phys. Rev. Lett. *97*, 038103.

Duhr, S., and Braun, D. (2006b). Why molecules move along a temperature gradient. Proc. Natl. Acad. Sci. U S A *103*, 19678–19682.

Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E., Guttman, M., and Lander, E.S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature *539*, 452–455.

Engstrom, P.G., Fredman, D., and Lenhard, B. (2008). Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. Genome Biol. *9*, R34.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

Groff, A.F., Sanchez-Gomez, D.B., Soruco, M.M., Gerhardinger, C., Barutcu, A.R., Li, E., Elcavage, L., Plana, O., Sanchez, L.V., Lee, J.C., et al. (2016). In vivo characterization of Linc-p21 reveals functional cis-regulatory DNA elements. Cell Rep. *16*, 2178–2186.

GTEx Consortium. (2013). The genotype-tissue expression (GTEx) project. Nat. Genet. *45*, 580–585.

Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large non-coding RNAs. Nature *482*, 339–346.

Harmston, N., Baresic, A., and Lenhard, B. (2013). The mystery of extreme non-coding conservation. Philos. Trans. R. Soc. Lond. B Biol. Sci. *368*, 20130021.

He, S., Liu, S., and Zhu, H. (2011). The sequence, structure and evolutionary features of HOTAIR in mammals. BMC Evol. Biol. *11*, 102.

Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P., and Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep. *11*, 1110–1122.

Hon, C.C., Ramilowski, J.A., Harshbarger, J., Bertin, N., Rackham, O.J., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T.M., Severin, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5′ ends. Nature *543*, 199–204.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. *47*, 199–208.

Kaikkonen, M.U., and Adelman, K. (2018). Emerging roles of non-coding RNA transcription. Trends Biochem. Sci. *43*, 654–667.

Lee, A.P., Koh, E.G., Tay, A., Brenner, S., and Venkatesh, B. (2006). Highly conserved syntenic blocks at the vertebrate Hox loci and conserved regulatory elements within and outside Hox gene clusters. Proc. Natl. Acad. Sci. U S A *103*, 6994–6999.

Li, L., Liu, B., Wapinski, O.L., Tsai, M.C., Qu, K., Zhang, J., Carlson, J.C., Lin, M., Fang, F., Gupta, R.A., et al. (2013). Targeted disruption of Hotair leads to homeotic transformation and gene derepression. Cell Rep. *5*, 3–12.

Li, L., Helms, J.A., and Chang, H.Y. (2016a). Comment on "hotair is dispensable for mouse development". PLoS Genet. *12*, e1006406.

Li, W., Notani, D., and Rosenfeld, M.G. (2016b). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nat. Rev. Genet. *17*, 207–223.

McEwen, G.K., Woolfe, A., Goode, D., Vavouri, T., Callaway, H., and Elgar, G. (2006). Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Res. *16*, 451–465.

Mercer, T.R., and Mattick, J.S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. Nat. Struct. Mol. Biol. *20*, 300–307.

Moon, M.H., Hilimire, T.A., Sanders, A.M., and Schneekloth, J.S., Jr. (2018). Measuring RNA-ligand interactions with microscale thermophoresis. Biochemistry *57*, 4638–4643.

Nobrega, M.A., Ovcharenko, I., Afzal, V., and Rubin, E.M. (2003). Scanning human gene deserts for long-range enhancers. Science *302*, 413.

Orom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al. (2010). Long noncoding RNAs with enhancer-like function in human cells. Cell *143*, 46–58.

Osterwalder, M., Barozzi, I., Tissieres, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. Nature *554*, 239–243.

Paralkar, V.R., Taborda, C.C., Huang, P., Yao, Y., Kossenkov, A.V., Prasad, R., Luan, J., Davies, J.O., Hughes, J.R., Hardison, R.C., et al. (2016). Unlinking an lncRNA from its associated cis element. Mol. Cell *62*, 104–110.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499–502.

Pereira, B., Chin, S.F., Rueda, O.M., Vollan, H.K., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.J., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat. Commun. *7*, 11479.

Portoso, M., Ragazzini, R., Brencic, Z., Moiani, A., Michaud, A., Vassilev, I., Wassef, M., Servant, N., Sargueil, B., and Margueron, R. (2017). PRC2 is dispensable for HOTAIR-mediated transcriptional repression. EMBO J. *36*, 981–994.

Pundhir, S., Bagger, F.O., Lauridsen, F.B., Rapin, N., and Porse, B.T. (2016). Peak-valley-peak pattern of histone modifications delineates active regulatory elements and their directionality. Nucleic Acids Res. *44*, 4037–4051.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Sandelin, A., Bailey, P., Bruce, S., Engstrom, P.G., Klos, J.M., Wasserman, W.W., Ericson, J., and Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics *5*, 99.

Schorderet, P., and Duboule, D. (2011). Structural and functional differences in the long non-coding RNA hotair in mouse and human. PLoS Genet. *7*, e1002071.

Selleri, L., Bartolomei, M.S., Bickmore, W.A., He, L., Stubbs, L., Reik, W., and Barsh, G.S. (2016). A Hox-embedded long noncoding RNA: is it all hot air? PLoS Genet. *12*, e1006485.

Sigova, A.A., Abraham, B.J., Ji, X., Molinie, B., Hannett, N.M., Guo, Y.E., Jangi, M., Giallourakis, C.C., Sharp, P.A., and Young, R.A. (2015). Transcription factor trapping by RNA in gene regulatory elements. Science *350*, 978–981.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al. (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. *3*, e7.

Yin, Y., Yan, P., Lu, J., Song, G., Zhu, Y., Li, Z., Zhao, Y., Shen, B., Huang, X., Zhu, H., et al. (2015). Opposing roles for the lncRNA Haunt and its genomic locus in regulating HOXA gene activation during embryonic stem cell differentiation. Cell Stem Cell *16*, 504–516.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature *515*, 355–364.

iScience, Volume *23*

Supplemental Information

Ancestrally Duplicated Conserved Noncoding Element Suggests Dual Regulatory Roles of HOTAIR in *cis* and *trans*

Chirag Nepal, Andrzej Taranta, Yavor Hadzhiev, Sachin Pundhir, Piotr Mydel, Boris Lenhard, Ferenc Müller, and Jesper B. Andersen

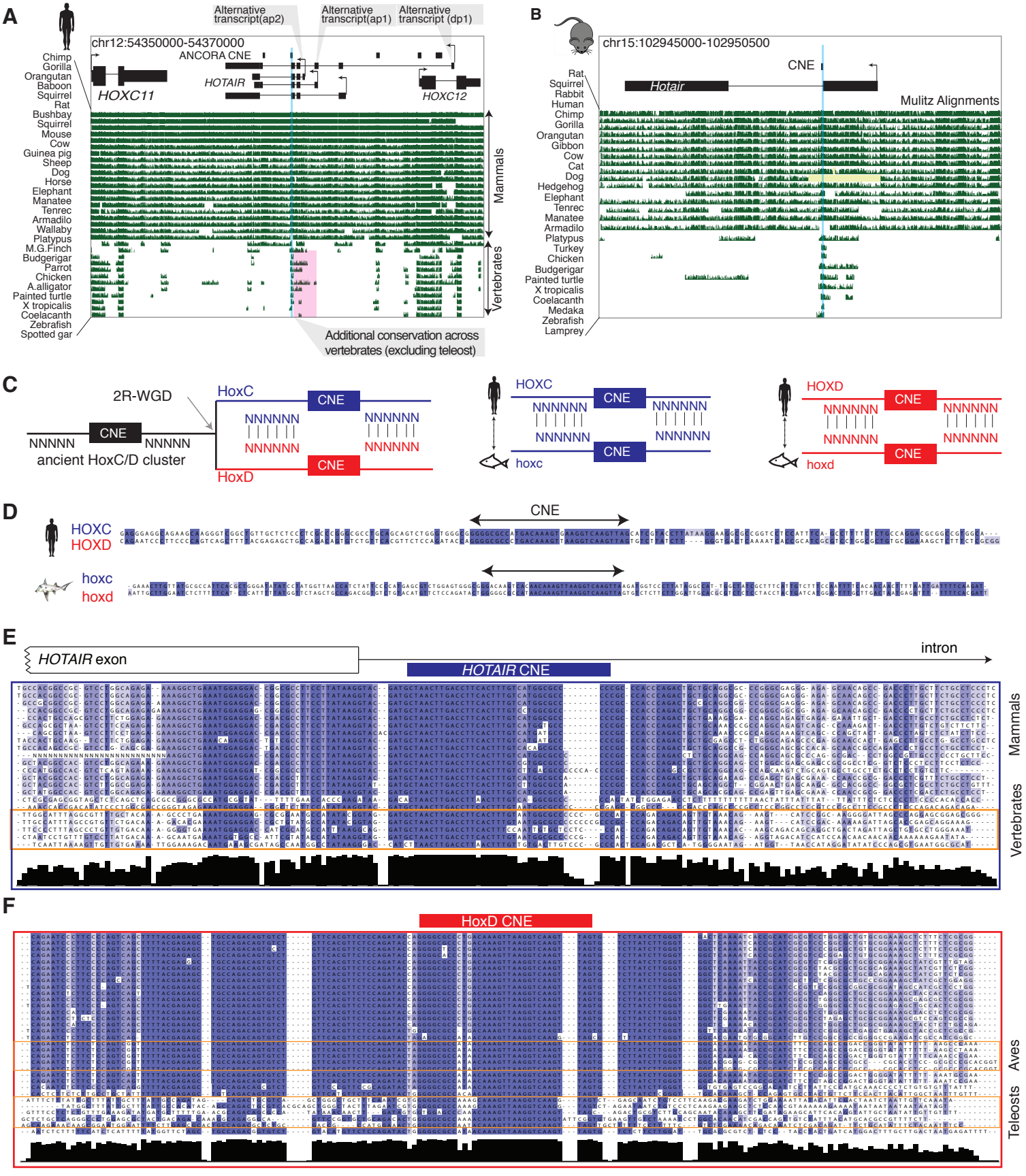


Figure S1. Sequence alignment of *HOTAIR*. Related to Figure 1. (**A-B**) A genome browser view of *HOTAIR* in human (A) and mouse (B) show sequence alignments across vertebrates. Conserved noncoding elements (CNEs) track from ANCORA browser is shown as on top. CNE that is conserved across all mammals and vertebrates is highlighted. (**C**) Schematic representation to show the origin of HoxC and HoxD cluster from the ancestral HoxC/D cluster after second round of whole genome duplication (2R-WGD). Schematic representations of HoxC and HoxD clusters separately across vertebrates. (**D**) Alignment of sequences flanking paralogous CNEs in human and elephant shark show little conservation despite both sequences being duplicated from the same ancestral sequences. (**E-F**) Alignment of sequences flanking *HOTAIR* CNE (**E**) and HOXD CNE (**F**). Species are aligned in the same order as in A.

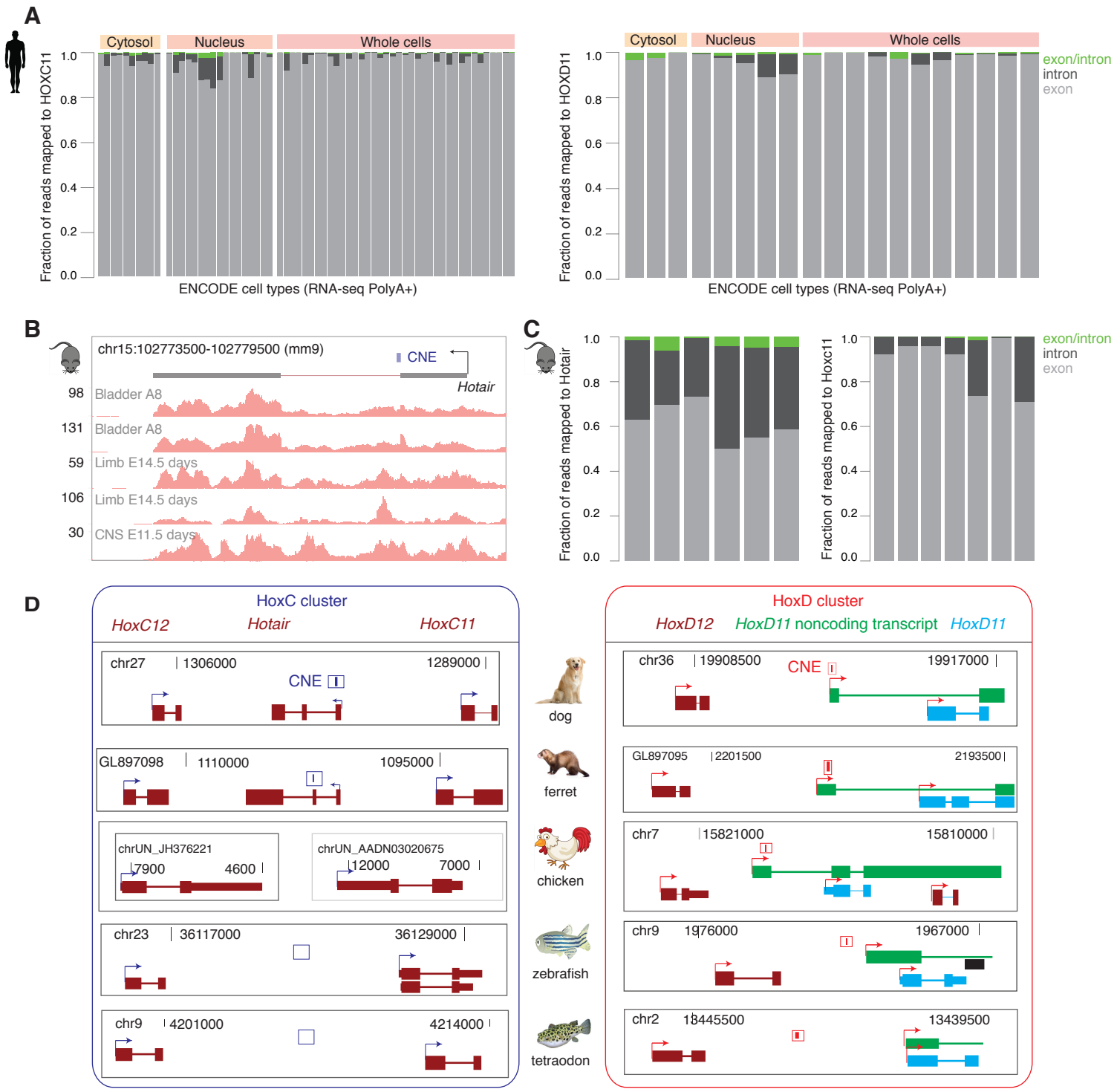


Figure S2. Paralogous CNEs are embedded in mature noncoding transcrips. Related to Figure 2. (**A**) Distribution of reads mapped to HOXC11 and HOXD11 exon, intron and overalpping exon/intron junctions across multiple cell types. Cells types are ordered based on increasing number of mapped reads. (**B**) A geneome browser view of Hotair transcript along with RNA-seq coverage tracks across different cell types from mouse ENCODE. (**C**) Distribution of reads mapped to mouse Hotair and Hoxc11 across multiple cell types. Cells types are ordered based on increasing number of mapped reads. (**D**) Evidence of *HOTAIR* and *HoxD11* noncoding transcript across multiple species. The CNEs are represented by rectangular blue and red bar in HoxC and HoxD cluster respectively. The *hoxc11* and *hoxc12* genes are assembled in different contigs in chicken, and homolog of *HOTAIR* CNE is undetected because the intergenic region between *hoxc11* and *hoxc12* is not assembled. Zebrafish *hoxc11* and *hoxc12* is assembled but lacks the CNE. *HoxD11* noncoding transcript is detected across tetrapods but not in teleosts (zebrafish and tetraodon).

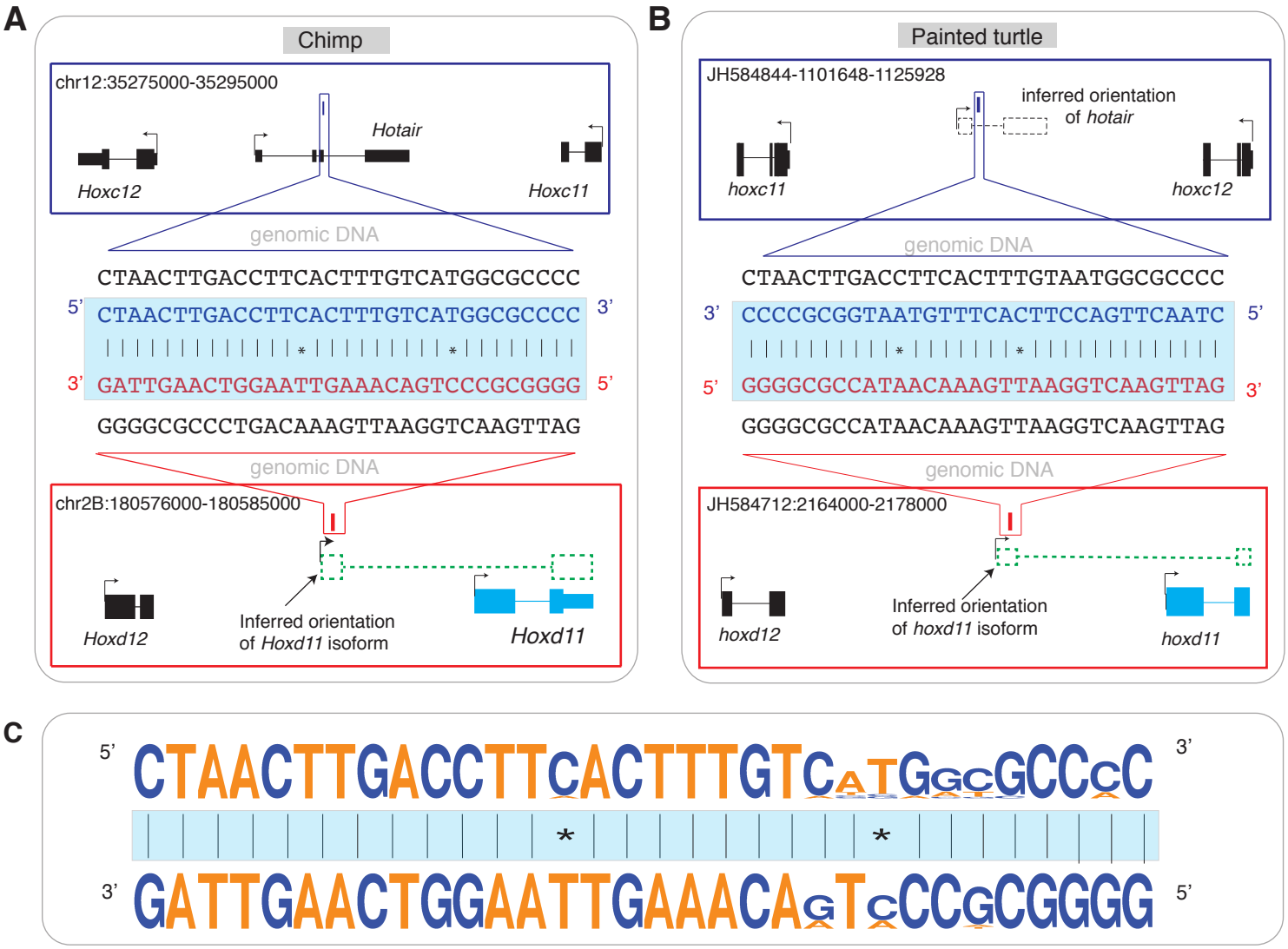


Figure S3. Paralogous CNEs exhibit sequence complementarity with respect to transcription directionality of *HOTAIR* and *HoxD11* noncoding transcript (*ncHOXD11*). Related to Figure 3. (**A-B**) Schematic representation to depict the inferred orientation of missing transcripts in chimp and painted turtle. *HOTAIR* is antisense to *HoxC11* gene, so the same convention was used to infer the orientation of *HOTAIR*. The *ncHOXD11* is an alternative splice variant of *HoxD11* coding gene across multiple species; thus, the same convention was used. Expected transcripts are represented by dashed rectangular boxes and lines. Arrows indicate directionality of transcript. The CNE sequences are zoomed in and shown as genomic DNA and transcribed RNA. Paralogogous CNEs exhibit sequence complementarity when aligned in 5' to 3' orientation. (**C**) Sequence logos of *HOTAIR* CNE and HoxD CNE show paralogous CNEs exhibit sequence complementarity in transcribed orientation.

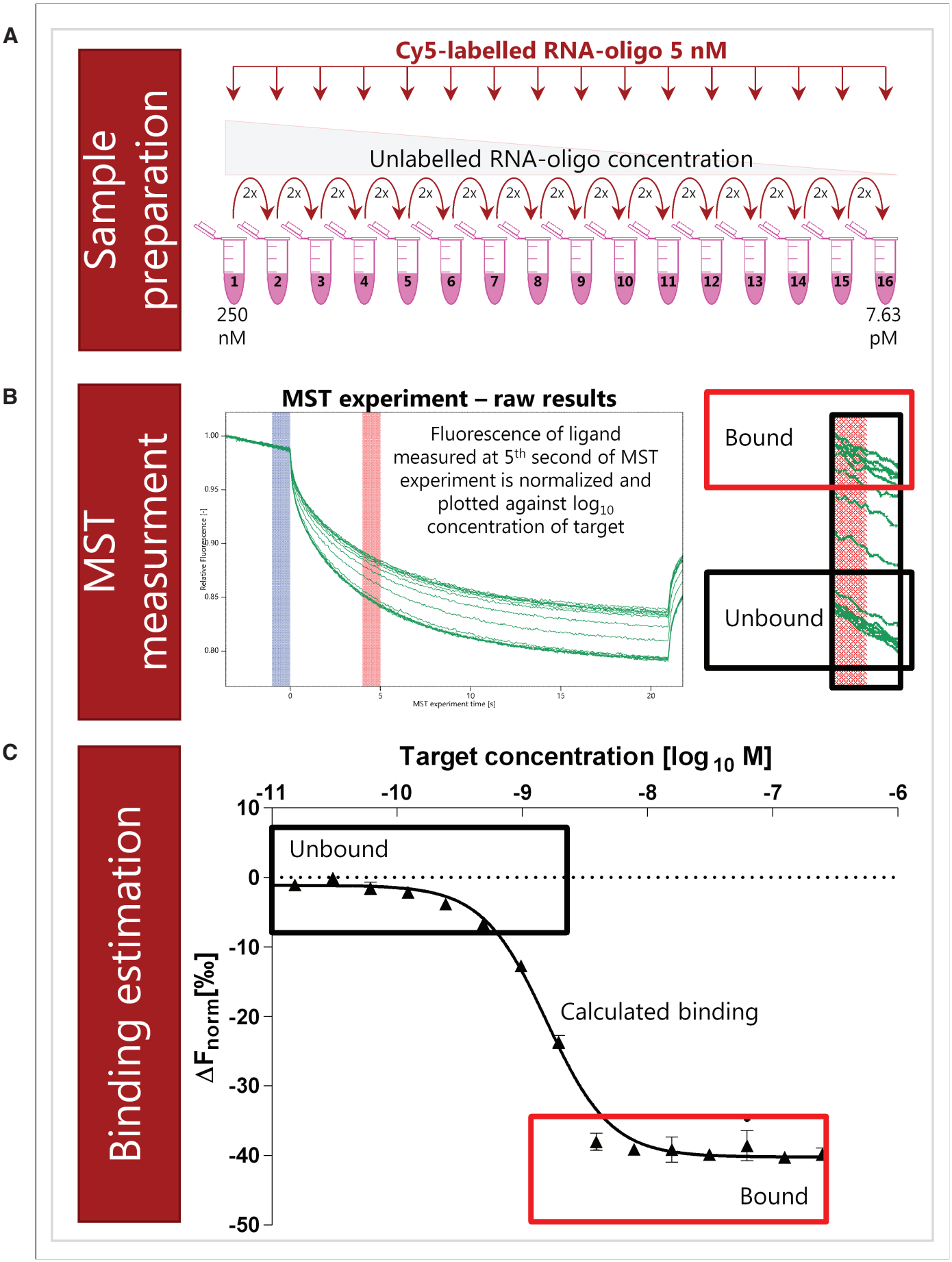


Figure S4. A schematic workflow to describe samples preparation for miscroscale thermophoresis (MST). Related to Figure 3. (**A**) Two-fold dilutions of unlabeled RNA-oligo were prepared starting at 250nM concentration. Labelled RNA-oligo was kept constant at 5nM. (**B**) An illustrative example of raw experimental data. Fluorescent of labelled RNA-oligo was measured at 5th second of the MST experiment. (**C**) Raw data were normalized as ΔFnorm [‰] and plotted against $\log_{10}$ concentration of titrated RNA-oligo.

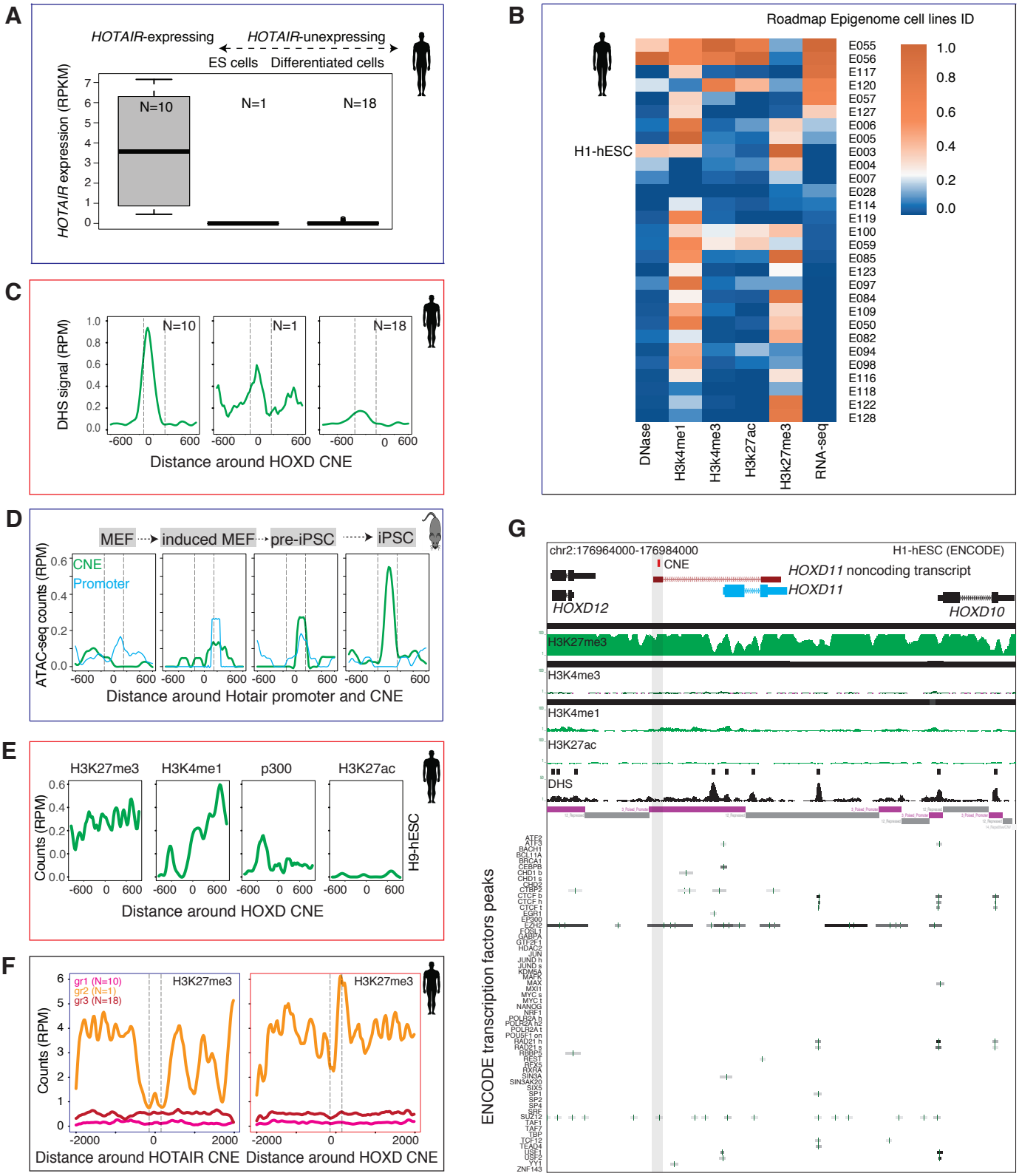


Figure S5. Transcription and chromatin environment of parlogous CNEs. Related to Figure 4. (**A**) Expression levels of *HOTAIR* in twenty-nine cell lines from Roadmap Epigenome data. A threshold of 0.5 RPKM (reads per kilobase per million mapped reads) was used to separate HOTAIR-expressing cell lines. HOTAIR-unexpressing cell lines were further separated into stem cells (N=1) and terminally differentiated cells (N=18). (**B**) Heatmap shows the normalized read counts in 250 nucleotides flanking *HOTAIR* CNE across 29 cell lines. (**C**) DHS signals around the HOXD CNE across three groups. Y-axis represents normalized counts in reads per million (RPM). (**D**) Dynamics regulation of chromatin state around mouse *Hotair* CNE during reprogramming of mouse embryonic fibroblast to iPSC. (**E**) Distribution of H3K4me1, H3K27me3, H3K27ac and p300 signals around human HOXD CNE in H9-hESC cell line. (**F**) Pattern of H3K27me3 marks around paralogous CNEs across three groups. (**G**) A genome browser view around HOXD CNE shows enrichment of multiple signals (transcription factors, DHS, histone modifications, chromHMM marks) in H1-hESC cell line.

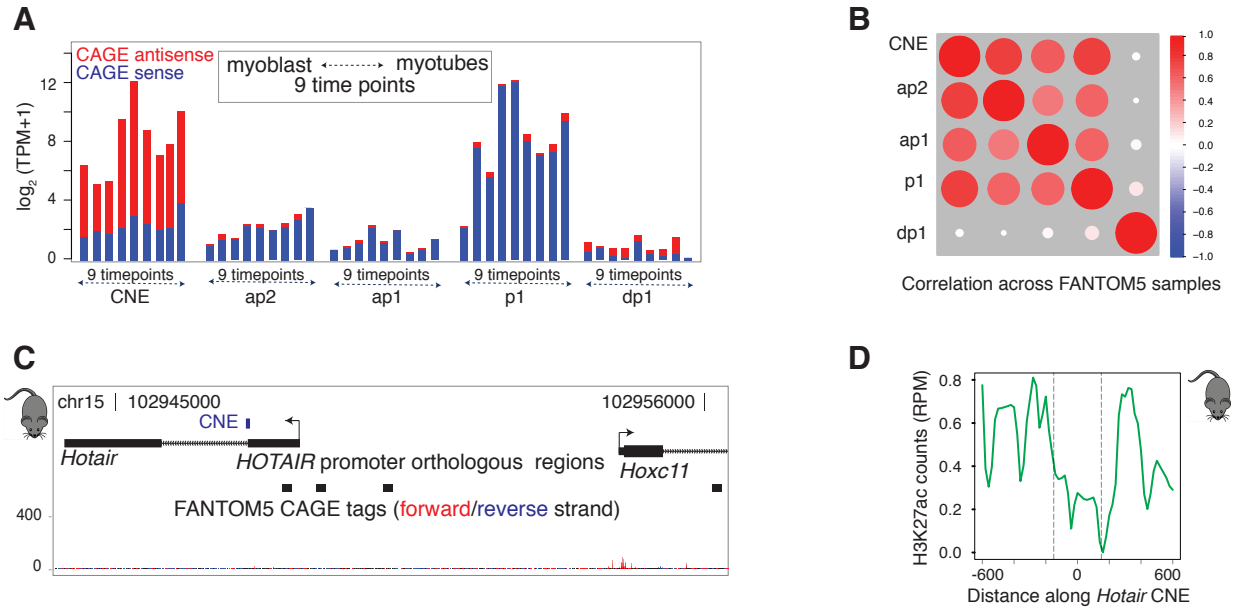


Figure S6. Transcriptional dynamics of *HOTAIR* promoters. Related to Figure 5. (**A**) Expression levels of *HOTAIR* promoter (p1), alternative promoters (ap1, ap2), distal promtoer (dp1) and CNE across nine different time points during differentiation from myoblast to myotube. (**B**) Expression correlation of *HOTAIR* promoter, alternative promoters and CNE across FANTOM5 samples are positively correlated except for distal promoter (dp1). (**C**) Distribution of CAGE tags around mouse *Hotair* locus in FANTOM5 samples. No significant CAGE tags are detected as it lacks tissues (embryonic hindlimbs, genital tubercle and a piece of trunk corresponding to sacro-caudal region) where *Hotair* is expressed. Orthologs of promoter region of *HOTAIR* are aligned to mouse. (**D**) *Hotair* CNE is flanked by bidirectional H3K27ac on mouse hindlimbs (E10.5 days).

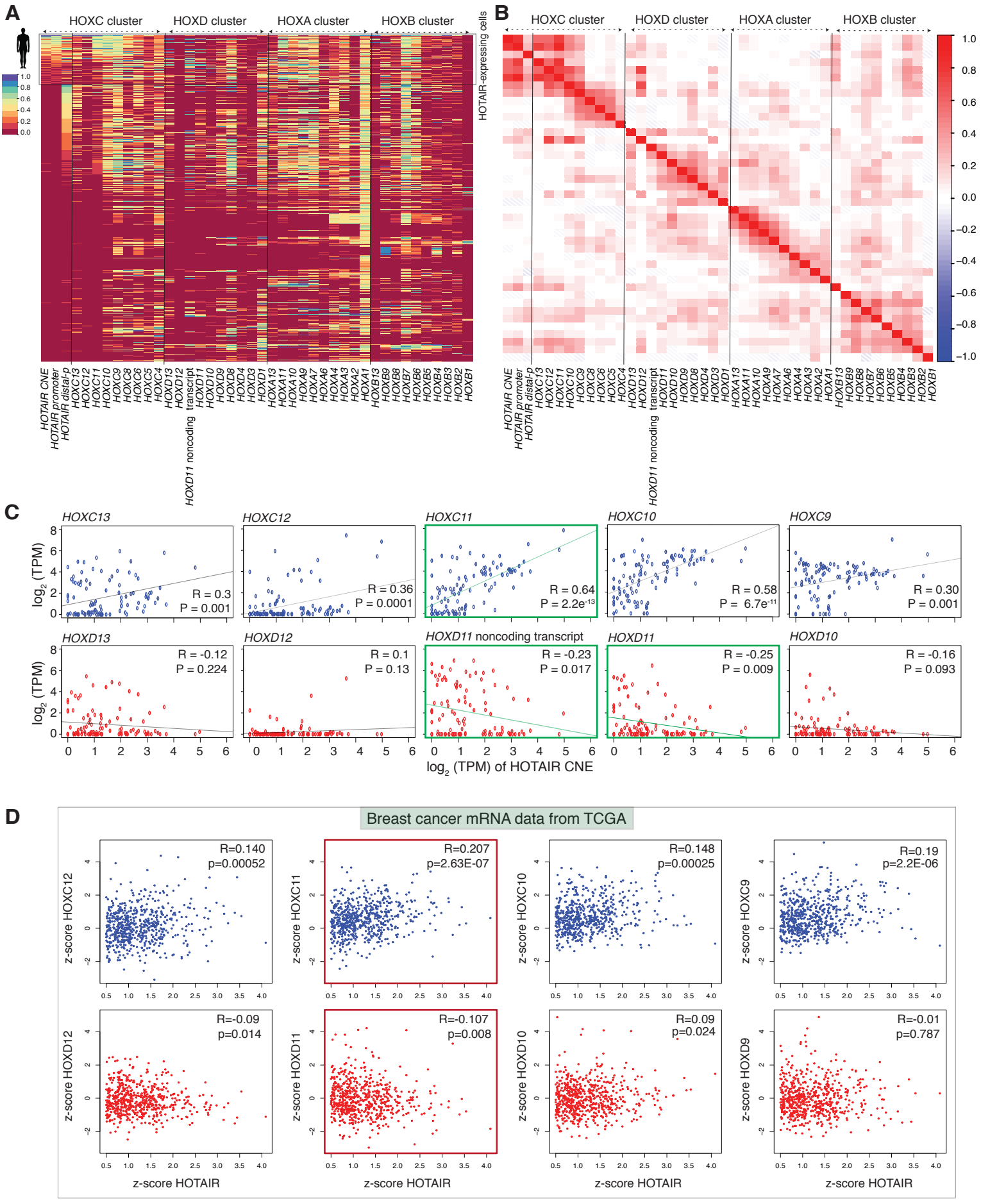


Figure S7. Co-expression analysis of *HOTAIR* with genes from four HOX clusters. Related to Figure 6. (**A**) Heatmap shows the expression level of *HOTAIR* and four HOX cluster genes across FANTOM5 samples. Cell types are sorted based on the expression levels of *HOTAIR* CNE. Expression levels are scaled between 0-1 across each column. (**B**) Correaltion of expression levels of *HOTAIR* with genes from four HOX clusters. (**C**) Correaltion of expression levels of *HOTAIR* CNE with HOXC and HOXD cluster posterior genes. Expression level is measured as tags per million (TPM). The most positively correalted gene (*HOXC11*) and most negatively correalted gene (*HOXD11* noncoding transcript) is highlighted in green border. **(D)** Correaltion of expression levels of *HOTAIR* with HOXC and HOXD cluster posterior genes on individual breast cancer patients.

Transparent Methods

Genome assemblies and gene annotations

Analyses on human and mouse were done in hg19 and mm9 genome version respectively. The genome assemblies of 37 species are listed in Table S1**.** Gene models were downloaded from UCSC (Speir et al., 2016). The conserved noncoding elements (CNEs) were downloaded from ANCORA (Engstrom et al., 2008).

Roadmap Epigenome data sets

Roadmap Epigenome data were downloaded from NIH Roadmap Epigenome browser (Roadmap Epigenomics et al., 2015). Annotated chromatin states were downloaded from 127 cell lines. Histone modifications (H3K4me1/3, H3K27ac/me3) and DNase I hypersensitive sites (DHSs) data were downloaded as mapped (tagAlign format) files. RNA-seq data for 57 cell lines were downloaded in the computed gene expression (RPKM) matrix (Roadmap Epigenomics et al., 2015). Only 29 cell lines that had all four (H3K4me1/3, H3K27ac/me3) histone modifications, DHS and RNA-seq were used for downstream analyses.

ENCODE and mouse ENCODE data sets

Histone modification and RNA-seq data from ENCODE were downloaded as mapped BAM files. ENCODE transcription factor ChIP-seq were downloaded as annotated peaks (Gerstein et al., 2012). Histone modifications (H3K4me1/3, H3K27ac/me3), DHS and RNA-seq data from mouse ENCODE were downloaded as mapped BAM files (Yue et al., 2014). Samples with replicates were merged into a single file. A threshold of 0.5 RPKM (reads per kilobase per million) was used as the cutoff expression to determine whether *HOTAIR* is expressed or not in the given RNA-seq samples.

FANTOM5 data sets

Human and mouse CAGE-seq data were downloaded from FANTOM5 (Arner et al., 2015; Consortium et al., 2014). Replicates were pooled into single file and resulting CAGE tags in each sample were quantified as tags per million (TPM). CAGE tags with the highest expression level were defined as the dominant transcription start site (TSS). CAGE based expression level was computed by summing all CAGE tags in the defined promoter region (300 bases upstream and downstream of TSS). To compare the expression correlation across four HOX clusters genes, we selected only that samples/cell types if any of HOX genes had a minimum expression level of 5 TPM, which resulted in a total of 694 cell types.

GTEx RNA-seq data sets

Mapped GTEx RNA-seq expression data (Consortium, 2013) for genes and transcripts were downloaded from GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz and GTEx_Analysis_2017-06-05_v8_RSEMv1.3.0_transcript_tpm.gct.gz respectively. These data contain 17382 samples from different tissues. For comparative analysis of expression levels of *HOXD11* coding and noncoding transcripts, we selected only those samples where both

transcripts had a minimum expression level of 0.1 TPM and additionally one of the transcripts had a minimum expression level of 0.5 TPM, which resulted in 1830 samples. For comparative analysis of expression levels of *HOTAIR*, *HOXC11* and *HOXD11* gene, we selected only those samples where all three transcripts had a minimum expression level of 0.1 TPM and additionally one of the transcripts had a minimum expression level of 0.5 TPM, which resulted in 2633 samples.

Data sets used from multiple studies

RNA-seq transcripts for multiple species were used from previous studies (Basu et al., 2016; Hezroni et al., 2015; Nepal et al., 2013). Raw data during reprogramming of mouse embryonic fibroblast to iPSC were download from GEO (GSE90894) (Chronis et al., 2017). Raw data for H3K27me3, H3K4me1, H3K27ac and p300 from H9-hESC cell lines were download from GEO (GSE24447) (Rada-Iglesias et al., 2011). Mouse embryonic (10.5 days) hind limb data were downloaded from GEO (GSE84793) (Andrey et al., 2017). Raw fastq reads were mapped using bowtie2 (Langmead and Salzberg, 2012). Only unique mapping reads were considered for downstream analysis. Breast cancer patients' mRNA (Illumina Human v3 microarray) data (Pereira et al., 2016) were downloaded from TCGA portal. Expression levels are measured in z-scores. We filtered samples where *HOTAIR* expression (<= 0.5) and were left with 605 patients.

Intron retention reads of *HOTAIR*

For intron retention analysis, we downloaded long RNA-seq data from ENCODE in human and mouse, in the form of mapped BAM files. For intron retention analysis, only polyA+ libraries were analyzed, and further classified into whole cell, nuclear fraction and cytosol fraction enriched libraries. To compute the ratio of intron and exon reads, we used gene annotation from RefSeq and computed the number of reads mapped to exons and introns. We only included samples if the total number of reads mapped to *HOTAIR* was higher than hundred. The sequence reads that were unspliced and overlapped the exon/intron junctions were counted separately from exonic and intronic reads.

Mapping of *HOTAIR* CNE across multiple species

The *HOTAIR* CNE sequences from both human and zebrafish was used as a query sequence. We used BLAST (blastall –p blastn –d –e 0.01 –m 8)(Altschul et al., 1997) to find homologous sequences against 37 species (Supplementary Table S1). Even at the permissive e-value cutoff of 0.01, only two homologous sequences were identified.

Annotation and directionality of *HOTAIR* and *HOXD11* noncoding transcripts overlapping CNEs

The *HOTAIR* transcript is annotated in multiple species (Hezroni et al., 2015; Speir et al., 2016), such as human, chimp, mouse, ferret and dog, and its orientation is antisense to *HoxC11* and *HoxC12* genes. Species lacking *HOTAIR* annotation, orientation of *HOTAIR*

CNE was assigned antisense to annotated HoxC cluster genes. In multiple species, such as human, chimp, mouse, ferret, dog and chicken, HoxD CNE is embedded within the exon of *ncHoxD11*, which is an alternative transcript of *HoxD11* coding gene. Thus, orientation of HoxD CNE was assigned similar to annotated *HoxD11* gene. Among teleosts fish, we analyzed RNA-seq transcripts in zebrafish (Hezroni et al., 2015; Nepal et al., 2013) and tetraodon (Basu et al., 2016), and did not identify *ncHoxD11*.

Software and tools

Multiple alignments were generated using ClustalW (Chenna et al., 2003) and Jalview (Waterhouse et al., 2009). Sequence logos were generated using WebLogo (Crooks et al., 2004). Data were visualized by uploading bigwig tracks on UCSC genome browser and images were downloaded. Bedtools (Quinlan and Hall, 2010), bash, perl and R scripts were used for data analysis.

Microscale thermophoresis experiment

The microscale thermophoresis (MST) is based on the phenomenon of molecule drift in temperature gradient (Asmari et al., 2018; Duhr and Braun, 2006a, b; Moon et al., 2018). In constant buffer conditions, thermophoresis depends on molecule size, charge and solvation entropy (hydration shell) which may change upon ligand binding. To measure the thermophoretic effect, the ligand is fluorescently labelled and kept at a constant concentration, whereas its interactor is titrated. Change in fluorescence emission at different ligand concentrations reflects an altered response based on the force of a temperature gradient. Plotting of fluorescent signal change against altered ligand concentration allow the calculation of Kd/EC_{50}.

Cy5-labelled or unlabelled RNA oligonucleotides (Supplementary Table S3) corresponding to CNEs and short flanks were used (TAG Copenhagen). To minimize potential influence of labelling on CNEs' interaction experiment was set up with mixtures of HOXD with labelled HOTAIR-Cy5, and HOXC with Cy5-HOXD. Fluorescent and regular RNA oligos dilutions were prepared in 1x MST buffer. In initial experiment MST buffer and MST buffer with addition of unspecific RNA was tested returning comparable results. Thus, data presented here were recorded on samples prepared in 1x MST buffer only. Unlabelled oligo was prepared as serial 2x dilutions in 15 µL volume accordingly to manufacturer recommendation. Fifteen µL of labelled 10 nM oligo was added to serial dilutions and mixed by pipetting. Final concentration of labelled RNA-oligo was 5nM and ligand was in range of 250 nM to 7.63 pM. Regular RNA-oligo corresponding to the labelled probe used in particular experimental setup was used as competitor. Details on samples and RNA-oligo types and concentration used are described (Supplementary Table S4). After short incubation, prepared mixtures were loaded into Standard Treated Capillaries and MST signal was measured on Monolith NT.115 (NanoTemper Technologies) with default settings (auto-detect LED and medium MST power).

Each experimental condition was run at least 3 times. Representative graphs of raw MST data are depicted on (Figure S4). For the analysis baseline corrected normalized fluoresce (ΔF_{Norm}) was used as recommended in MST software manual, and plotted against the log10 ligand concentration in GrapPad Prism 7 (GrapPad Software). The threshold values were extrapolated from sigmoidal fitting curve.

Supplemental References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Andrey, G., Schopflin, R., Jerkovic, I., Heinrich, V., Ibrahim, D.M., Paliou, C., Hochradel, M., Timmermann, B., Haas, S., Vingron, M.*, et al.* (2017). Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. Genome Res *27*, 223-233.

Arner, E., Daub, C.O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drablos, F., Lennartsson, A., Ronnerblad, M., Hrydziuszko, O., Vitezic, M.*, et al.* (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science *347*, 1010-1014.

Asmari, M., Ratih, R., Alhazmi, H.A., and El Deeb, S. (2018). Thermophoresis for characterizing biomolecular interaction. Methods *146*, 107-119.

Basu, S., Hadzhiev, Y., Petrosino, G., Nepal, C., Gehrig, J., Armant, O., Ferg, M., Strahle, U., Sanges, R., and Muller, F. (2016). The Tetraodon nigroviridis reference transcriptome: developmental transition, length retention and microsynteny of long non-coding RNAs in a compact vertebrate genome. Sci Rep *6*, 33210.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. (2003). Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res *31*, 3497-3500.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. Cell *168*, 442-459 e420.

Consortium, F., the, R.P., Clst, Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberle, V., Lassmann, T.*, et al.* (2014). A promoter-level mammalian expression atlas. Nature *507*, 462-470.

Consortium, G.T. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580-585.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

Duhr, S., and Braun, D. (2006a). Optothermal molecule trapping by opposing fluid flow with thermophoretic drift. Phys Rev Lett *97*, 038103.

Duhr, S., and Braun, D. (2006b). Why molecules move along a temperature gradient. Proc Natl Acad Sci U S A *103*, 19678-19682.

Engstrom, P.G., Fredman, D., and Lenhard, B. (2008). Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. Genome Biol *9*, R34.

Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R.*, et al.* (2012). Architecture of the human regulatory network derived from ENCODE data. Nature *489*, 91-100.

Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P., and Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep *11*, 1110-1122.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Moon, M.H., Hilimire, T.A., Sanders, A.M., and Schneekloth, J.S., Jr. (2018). Measuring RNA-Ligand Interactions with Microscale Thermophoresis. Biochemistry *57*, 4638-4643.

Nepal, C., Hadzhiev, Y., Previti, C., Haberle, V., Li, N., Takahashi, H., Suzuki, A.M., Sheng, Y., Abdelhamid, R.F., Anand, S.*, et al.* (2013). Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. Genome Res *23*, 1938-1950.

Pereira, B., Chin, S.F., Rueda, O.M., Vollan, H.K., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.J., *et al.* (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat Commun *7*, 11479.
Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.
Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279-283.
Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., *et al.* (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330.
Speir, M.L., Zweig, A.S., Rosenbloom, K.R., Raney, B.J., Paten, B., Nejad, P., Lee, B.T., Learned, K., Karolchik, D., Hinrichs, A.S., *et al.* (2016). The UCSC Genome Browser database: 2016 update. Nucleic Acids Res *44*, D717-725.
Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.
Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., *et al.* (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature *515*, 355-364.